POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Evaluation of Demand Forecast Models for Urban Carsharing

ELHAM KARIMI

Départment de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Avril 2019

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Evaluation of Demand Forecast Models for Urban Carsharing

présenté par **Elham KARIMI**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Jean-Marc FRAYRET, président

Martin TRÉPANIER, membre et directeur de recherche

Vahid PARTOVI NIA, membre et codirecteur de recherche

Francesco CIARI, membre

DEDICATION

This thesis is dedicated to

my father: Reza Karimi,

my mother: Mehri Pourmand,

and my husband: Mahmoud Ramin.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Martin Trépanier who contributed to my project from the beginning to the end. He inspired me with his kindly continuous financial and spiritual support, knowledge and passion. It would not have been possible without his guidance and constantly availability throughout my research.

I would like to express my gratitude to my co-supervisor, Dr.Vahid Partovi Nia, for his great ideas, knowledge and taking part in this work despite his busy schedule. It was a great pleasure to know him and learn from his invaluable suggestions throughout my whole studies.

Financial support of this work was provided by Communauto carsharing operators. I am grateful for giving me the opportunity to do my master's thesis. Without their trust this project would be impossible.

I am grateful to Dr. Mina Mirshahi for proofreading different parts of this thesis.

Most importantly, my deepest gratitude goes to my parents, Reza and Mehri, and my brothers for all their unconditional love, engagements, dedications and support from miles away.

Last but definitely not the least, I would like to express my deepest gratitude and appreciation to my lovely husband, Mahmoud, for unconditional love and support. I am indebted to him every bit of success. He is the most important motivation for me to move forward and reach to my dreams.

RÉSUMÉ

Au cours des dernières décennies, les services de mobilité partagée ont été créés en tant que nouvelles alternatives de transport urbain. Le système d'autopartage est l'un de ces services récents impliquant une flotte de véhicules dispersés dans une ville. Cela a permis de contrecarrer certains problèmes, tels que le stationnement limité dans les zones denses de la ville, la pollution, l'augmentation de la possession automobile, etc. Communauto est l'un des plus anciens services d'auto-partage en Amérique du Nord, établi depuis 1994 au Canada.

Le but de ce mémoire est d'appliquer les méthodes d'apprentissage automatique les plus courantes afin de prévoir les heures et les kilomètres consommés selon les réponses souhaitées sur le système opérateur de Communauto avec deux services urbains à Montréal : régulier et en libre-service. La combinaison de différents modèles statistiques et de réseaux de neurones artificiels a été évaluée par le biais d'un ensemble d'expériences afin de prévoir la demande à partir de données historiques. Les modèles appliqués mettent l'accent sur la mise en œuvre de régressions multiples, d'arbres de régression, de forêts aléatoires, de *gradient boosting*, et des réseaux neuronaux récurrents basés sur *long short-term memory* et les *gated recurrent unit*.

Les modèles ont été appliqués aux données de Communauto Montréal. Par la suite des données supplémentaires, telles que les informations sur les journées de vacances et les conditions météorologiques, ont été associées aux données de Communauto Montréal afin de déterminer si les performances des modèles de prévision sont améliorées. Il est à noter que les données relatives au service régulier et aux heures consommées, en tant que réponse, ont été prises en compte pour le processus de prévision.

Les modèles ont été évalués par l'erreur quadratique moyenne sous forme d'indice de mesure de distance entre les valeurs réelles et les valeurs prédites sur la base des ensembles de test. Les ensembles de tests ont été examinés séparément dans les délais suivants : 2012, 2013, 2014 et 2015 à 2016. La moyenne des résultats a ensuite été considérée comme l'erreur finale de chaque modèle.

Les résultats montrent que les modèles statistiques tels que l'intensification du gradient en service régulier par rapport au nombre d'heures consommées (avec un taux d'erreur = 1437,48) étaient supérieurs aux modèles de réseaux neuronaux artificiels (taux d'erreur de LSTM = 2159,05, taux d'erreur de GRU = 2215,14). De plus, des facteurs supplémentaires ont amélioré la capacité des modèles de prévision, le taux d'erreur de renforcement du gradient ayant été considérablement

réduit à 1211,96. De plus, les résultats des modèles de prévision en service flottant en ce qui concerne le nombre d'heures consommées et le kilométrage montrent que la régression multiple surpasse les modèles de réseau neuronal artificiel. En outre, les facteurs supplémentaires ont considérablement amélioré les performances des modèles appliqués.

ABSTRACT

In the last few decades, shared mobility services have been made as new urban transport alternatives. Carsharing system is one of these recent services that involves a fleet of scattered vehicles in a city. This helped to counteract some problems, such as limited parking within city in dense areas, pollution, increase in car ownership, etc. Communauto is one of the oldest carsharing services in North America which has been established since 1994 in Canada.

The focus of this thesis is to apply the most common machine learning methods in order to forecast consumed hours and kilometers driven or mileage as desired responses at Communauto operator system with two urban services in Montreal: regular and free-floating. Combination of different statistical and artificial neural network models were evaluated through a set of experiments in order to forecast demand from historical data. The applied models include multiple regression, regression tree, random forests, gradient boosting, long short-term memory and gated recurrent unit based recurrent neural networks.

The models were applied to the Montreal Communauto data. Thereafter, additional factors such as holiday information and weather condition were engaged to the Communauto data to explore whether the performance of forecasting models was enhanced. It is noteworthy that the data related to regular service and consumed hours, as response, were considered for the forecasting process.

The models were evaluated by root mean squared error as an index of distance measurement between real and predicted values based on test sets. The test sets were considered separately in the following time frames: 2012, 2013, 2014, and from 2015 to 2016. The average of results was then considered as the final error of each model.

The results show that statistical models such as gradient boosting in regular service with respect to consumed hours (with error rate= 1437.48) outperformed artificial neural network models (error rate of LSTM = 2159.05, error rate of GRUs = 2215.14). Moreover, additional factors improved the ability of the forecasting models, as the error rate of gradient boosting was significantly reduced to 1211.96. Furthermore, the results of forecasting models in free-floating service with respect to consumed hours and mileage show that multiple regression outperformed artificial neural network models. Besides, the additional factors significantly improved performance of the applied models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AIC	Akaike Information Criterion
ANNs	Artificial Neural Networks
AUM	Auto-Mobile Service
BIC	Bayesian Information Criterion
CSV	Comma Separated Values
GRUs	Gated Recurrent Units
IQR	Interquartile
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
Q	Quartile
REG	Regular Service
RMSE	Root Mean Squared Error
RNNs	Recurrent Neural Networks
Tanh	Hyperbolic Tangent

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

1.1 Background

Over the years, carsharing services as an alternative to private vehicles with the ability to share cars with other users have gained significant traction. Members of carsharing services have access to a fleet of scattered vehicles without owning them. In this sense, users can take advantage of using vehicles privately regardless of the concerns expressed by the lease payments, insurance, gas filling, maintenance or parking. Therefore, cost effectiveness and accessibility are the main incentives for joining as a user of carsharing operators. Carsharing services have been developed and they have turned into one of an important and efficient element in moving passengers in local and suburban areas (Murray, Davis, Stimson, & Ferreira, 1998). More importantly, carsharing program as a modern approach has had a massive impact on environmental and transportation issues (Katzev, 2003). Therefore, the program has been extremely prosperous worldwide. Considering this fact, a lot of agencies have been fascinated by this kind of transportation, showing that the number of carsharing programs has been expanding around the world (Shaheen & Cohen, 2007).

Communauto is a carsharing organization pioneer based in Montreal, Canada, which has been one of the oldest and fastest enlarging carsharing services since 1994 in North America. Communauto carsharing services have been available in several cities in Quebec province of Canada, including Montreal, Quebec City, Sherbrooke, and Gatineau. In addition, carsharing services have been supplied by this organization under the name of VRTUCAR in the province of Ontario and Nova Scotia. Moreover, it has recently been developed in capital of France, Paris, and the suburbs. Fleeting car rental which has been promoted by this organization is comprised of 2 packages: regular and free-floating services. Regular carsharing operation, sometimes called round-trip, is developed with station-based systems by-which cars should be returned to the same dedicated station where users pick up. Travelers are charged for each kilometer and hour on every journey. Free-floating operator which offers Auto-Mobile vehicles, is planned to be one-way and is dispersed aboard specified zones without stations. Hence, this model in contrast to regular service allows users to take vehicles and drop off cars at any convenient on-street space within the legal zone. The term of this service is based on per minute fare. In both services, registered users can make a request to reserve available cars online or through smartphone applications. However, for

free-floating, the reservation window is 30 minutes, which is much smaller than for the regular service, in-which cars can be reserved up to 30 days ahead. Figure 1.1-1 shows a distributed fleet of vehicles at different locations in Montreal.

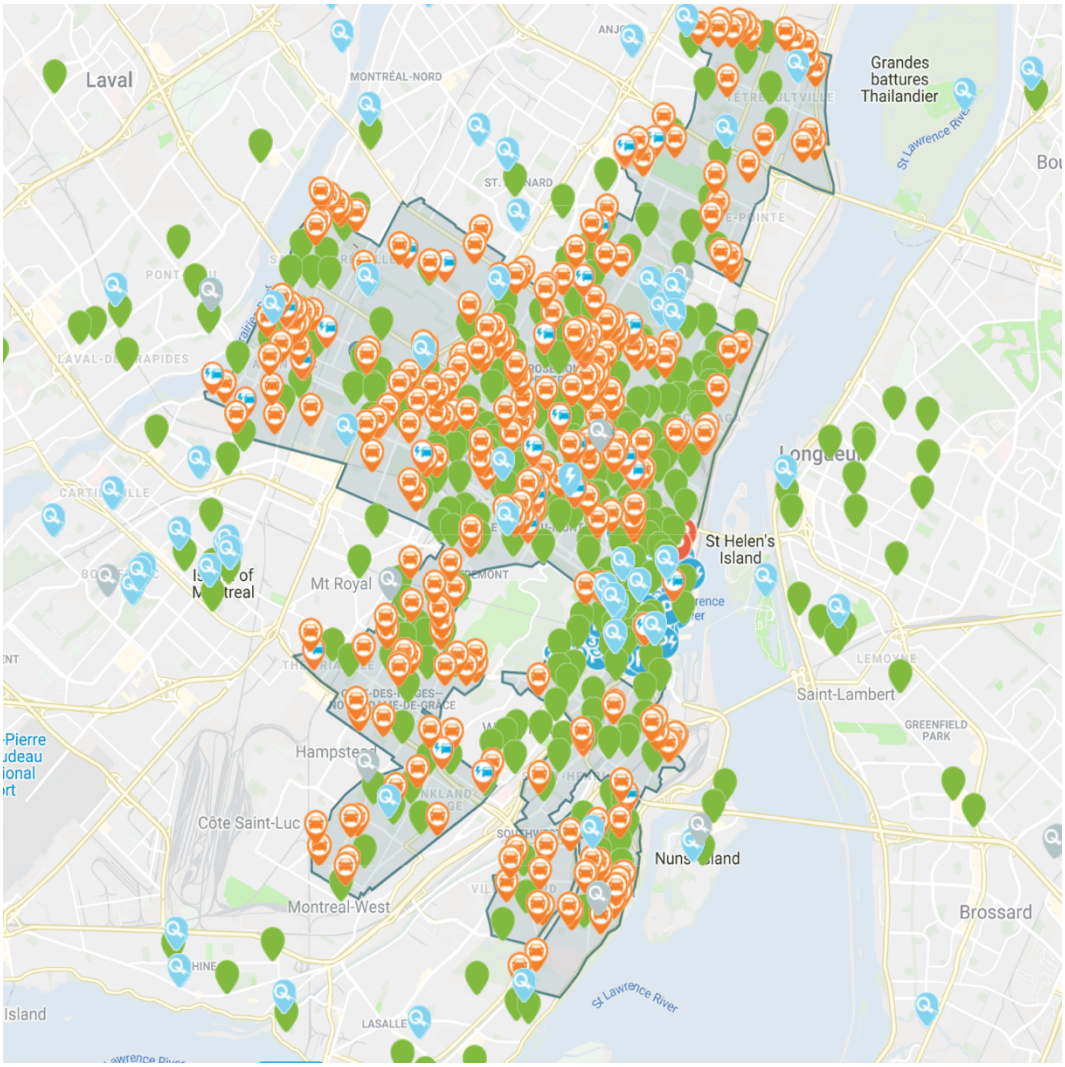

Figure 1-1: Communauto operating distribution in real-time status map in Montreal (Communauto Inc, 2019). The orange pins indicate available automobiles in free-floating service. The green pins show the active stations in regular service.

In the last few years, the sizeable applicants have become more mindful to join Communauto carsharing networks. Therefore, the ratio of members has been drastically rising. Due to changes in demand and impressive growth of members, carsharing operators have faced complex challenges

of restrictions on supply. In the light of this situation, the operation efficiency of this system relies on recognition of users' demands, and the benefit which will assist the company's survival for serving their services. Discovering about usage patterns must be an essential element to overcome restrictions and quick progress of the company.

Technology advancements have allowed generating data day-to-day. The collected data in this project are integrated by tracking the daily trips of users along with more details such as consumed hours and driven kilometers. Additionally, the agency's facilities include the number of active stations (for regular service) and available vehicles.

Machine learning methods are available in building accurate forecasting models and recognizing trends or patterns as they could emphasize a significant impact on the quality of the system. However, traditional methods may be more accurate in some applications, and it is relevant to examine different models to see their forecasting performance.

1.2 Research Project

The main focus of this thesis is developing methods aimed to forecast the demand of carsharing service, then assess the tuned models and their performance. The demand values of regular and free-floating services are defined as 1) consumed hours and 2) mileage (km). Indeed, the task is to forecast the future demand given series of historical observations, and moreover, investigating the factors affecting vehicle usage in Communauto carsharing operator, such as holiday and weather conditions, to be taken into optimizing the ability of the models.

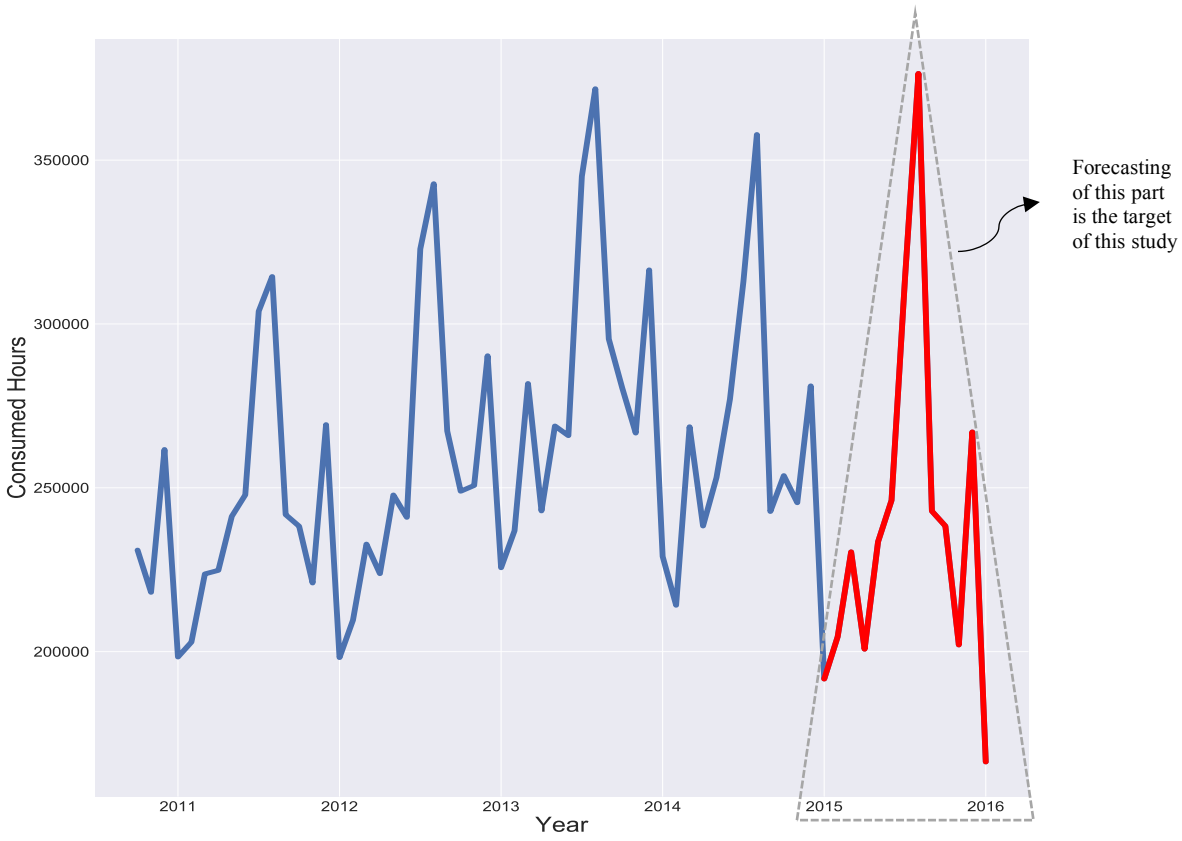

Figure 1-2: Example trend of consumed hours during a year in regular service of Communauto operation.

For example, Figure 1.2 shows the overall trend of consumed hours as one of the demands of this thesis. The blue colored observations are used to train the machine learning algorithms. The target is to forecast the red colored part.

1.3 Research Objectives

Apart from developing a demand forecasting model, the specific objectives of this work are to determine whether the forecasting power of the "traditional" statistical models outperform that of artificial neural networks models.

In order to explore the goal of the research, the following experiments are performed:

- Visualizing the dataset to detect data quality issues such as outliers, then perform cleaning.

- Exploring and seeking strong patterns and regularities between independent variables which are referred to features in this thesis.
- Applying forecasting models such as multiple regression, regression tree, random forests, gradient boosting, long short-term memory (LSTM) recurrent neural networks and gated recurrent units (GRUs) based on recurrent neural networks (RNNs) algorithms.
- Training the model on different samples of the data.
- Evaluating the models and comparing the models' performance

The forecasting models would be beneficial for the Communauto operation in order to forecast consumed hours and mileage as our dependent variables in this research. The results can undoubtedly assist the company to enhance further facilities to the existing members, thereby better expand their services in upcoming years. Moreover, this study will assist in better understanding of the demand for vehicles in the Communauto carsharing network.

1.4 Thesis Structure

The layout of the current study is organized as shown in Figure 1.3, with 5 chapters in total. Each chapter is separated into different sections:

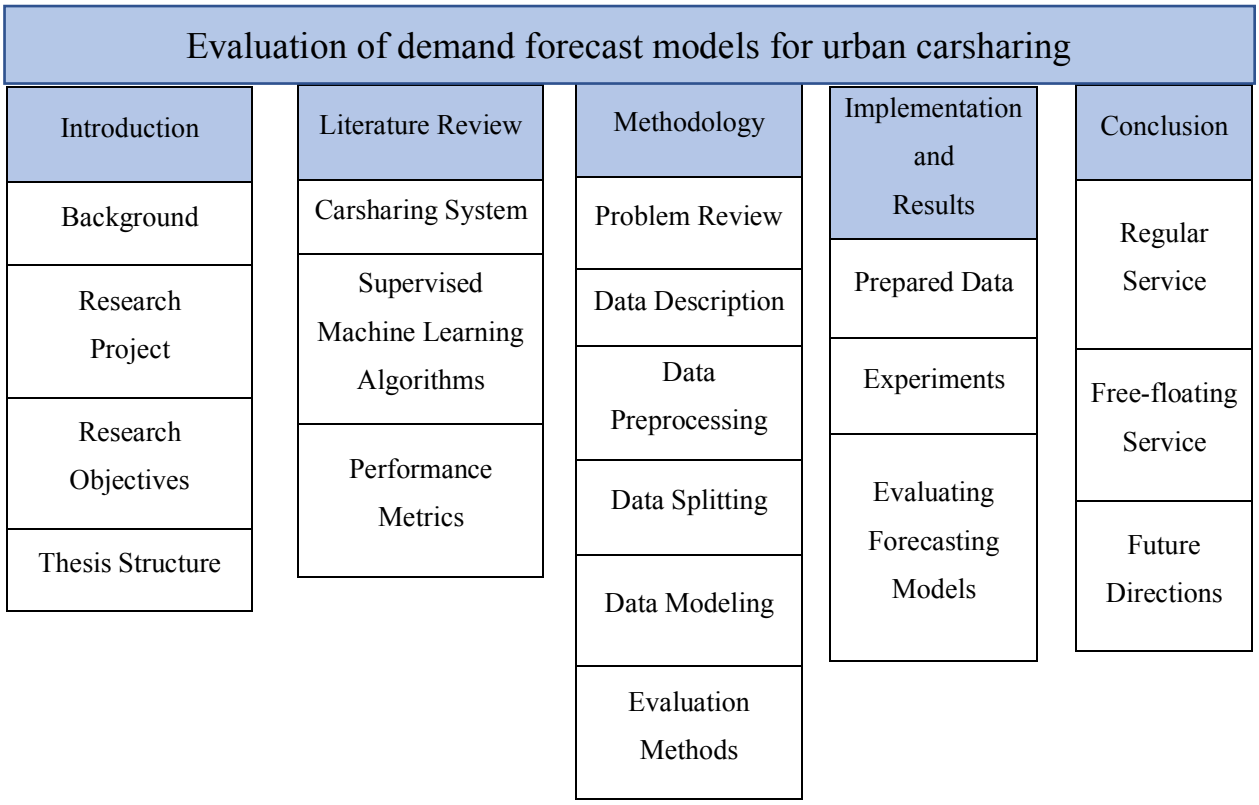

Figure 1-3: Thesis outline

- Chapter 2 (Literature Review) gives an overview of carsharing systems. Moreover, this chapter is conducted to reach an insight into machine-learning algorithms and the related work in supervised learning models such as multiple regression, regression tree, random forests, gradient boosting, LSTM and GRUs recurrent neural networks. Then, performance metrics are discussed in this part.
- Chapter 3 (Methodology) introduces several enriched datasets under this study. Then, relevant features, which contribute most to forecast outputs or responses in which this thesis is interested in, are extracted. Moreover, data pre-processing and visualization of data is done before applying the intended supervised techniques. Then, the methods which are used to create forecasting models are explained in detail. Finally, evaluation methods are introduced in this chapter.
- Chapter 4 (Results) implements the methods which are introduced in chapter 3 to clean the data from the noisy points and missing values. Afterward, the prepared data is used to learn the forecasting models such as multiple regression, regression tree, random forests, gradient

boosting, LSTM and GRUs recurrent neural networks. Finally, the performance of each model is evaluated using the root mean squared error (RMSE).

- Chapter 5 (Conclusion) discusses the results of the forecasting models on Communauto dataset with and without holiday and weather variables as additive features, as well as further research in the field.

CHAPTER 2 LITERATURE REVIEW

This chapter gives an insight provided on what the pertinent work is done in the field of carsharing systems and supervised machine learning algorithms. The review is widely arranged into three discussions. In section 2.1, the general background about carsharing services is described. Subsequently, the application and information about supervised machine learning models in forecasting are given in section 2.2. The discussion is regarding previous work done by modeling algorithms such as multiple regression, regression tree, random forests, gradient boosting, LSTM and GRUs recurrent neural networks which are ensemble machine learning models. In addition, section 2.3 provides the reflection of evaluating the models and method selection.

2.1 Carsharing System

While the world population has experienced rapid growth, transportation systems boast a remarkable role in economic satisfaction impacts, improve levels of life quality, however, causing excessive automotive mobility and threatening the environmental sustainability.

Shared mobility refers to transportation services that are shared among users which is nowadays used very broadly. It is including public transit, taxi, bike sharing, carsharing and other transportation modes.

Since the mid-1980s, carsharing systems as an easily accessible tool, have become popularized gradually across the world (Murray et al., 1998). The concept of carsharing is based on that the number of needed cars to provide the demand of members are dramatically less than when each user owns privet vehicles. Carsharing service provides users with access to shared vehicles for usually short-term use. The shared vehicles are scattered within a network of locations in a city. Carsharing creates the possibility to access a car at any time with a reservation, and the users are charged by either time or mileage. Round-trip and one-way are the most common models of carsharing operation. In both models, the fleet of vehicles are used for several trips by multiple users throughout the day. In round-trip, vehicles must be returned to the same location where users borrow the car. However, one-way system allows customers to take a vehicle at one location and drop it off at the other location.

This type of transportation mode helps to reduce car ownership with the growing access to use a shared fleet of vehicles which is more affordable than owning a personal car. Additionally, studies by Steininger and colleagues have shown that most European carsharing users do not own a car

(Steininger, Vogl, & Zettl, 1996). In this sense, users can take advantage of using shared vehicles regardless of the concerns about wasting time to rent a car from agencies and getting charged for the full day. In fact, users can reserve cars online or by smartphone applications prior to using them. Moreover, vehicles can be picked up at the nearest specified zones or stations with tolerable walking. Shared fleet organizations are responsible for paying all expenses of vehicle maintenance and repairs. Meanwhile, parking and insurance coverage are provided through them. So far, carsharing has had a significant effect on reducing vehicle ownership around the world. For instance, in 2001, B. Robert, the founder of Communauto in Canada, reported that 25% of the members of this organization sold their vehicles and more than 50% were able to avoid buying a car (Katzev, 2003).

In late 1994, Communauto was launched as one of the largest commercial organizations in North America, based in Montreal, Quebec. Most of its stations and zones are placed within residential areas, although some are located in central business areas or near transit nodes. Travellers are charged by the time and mileage which depends on the type of services that they use.

Obviously, Communauto carsharing service in Montreal has taken a major contributor to a mode of transportation along with low-pollutant emission vehicles. Moreover, Sioui et al. conducted a study on Communauto members which shows that most users who do not have private vehicles and often use carsharing services, drive 30 percent less than those who own private vehicles (Sioui, Morency, & Trépanier, 2013). Insofar as a study on carsharing systems is described by Martin and Shaheen, it presents a significant decrease in vehicle-kilometers driven after joining carsharing systems, although the drivers had their personal vehicles (Martin & Shaheen, 2011). Another positive effect which is assessed by the same researchers is the reduction of greenhouse gas emission in Canada and the US.

2.2 Supervised Machine Learning Algorithms

In the last 20 years, machine learning algorithms have been evolved as an important foundation of information technology. Machine learning algorithms are drastically applied with a variety of approaches in different problems which let computers exploit information by observing raw data and extracting patterns, then analyzing and optimizing performances (Witten, Frank, Hall, & Pal, 2016). This knowledge is remarkably applicable in high-level of computing (Y. Chen et al., 2014). Machine learning, a subfield of computer science, has generated a revolution in statistical science.

The expression of learning comes up with improvement of algorithms with respect to past experience and gained knowledge (Das & Behera, 2017).

In computer science, an algorithm is a set of instructions in order to express operations to find the most efficient way of carrying out response process of the object. In machine learning, algorithms are used to design mathematical models in order to produce informative results or uncover reliable associations hidden within observations (Alpaydin, 2009). Ongoing research and daily activities interact directly with machine learning, including weather forecasting, managing transportation, pattern recognition, fraud detection, medical diagnosis, and many other complex analytic purposes. Understanding the data is critical in order to recognize and implement various methodologies with more efficiency. In this setting, depending on existing data, most machine learning algorithms are broadly characterized into one of two categories: supervised learning and unsupervised learning. Supervised learning is focused on the samples where data has been labeled, and input and response are known. Indeed, this type of data allows a system to extract the hidden structure of data and builds the robust forecasting of unseen or test data (Mohri, Talwalkar, & Rostamizadeh, 2012). On the other hand, the data without labels are elaborated as unsupervised learning or clustering techniques. Whereas, in no ascribe labels on data, the learner detects the similarities between points, and then spreads them into various categories. Accordingly, each category takes the new label (Hastie, Tibshirani, & Friedman, 2009).

Some supervised learning models are proclaimed to overcome noisy data or overfitting issues to make a robust model. On the other hand, some models are interpretable and low flexible, while others are non-interpretable and highly flexible (Lipton, 2016). Practitioners should ask which models are matched to task and available dataset. In the following section, a detailed review of the concept and previous studies of applied models like multiple regression, regression tree, random forests, gradient boosting, RNNs with LSTM and GRU units are considered.

2.2.1 Multiple Regression

The initial form of regression (least squares) was published in early nineteenth century by Legendre and Guass, and the term of regression was later expressed by Francis Galton (Bingham, 2006). Regression as a statistical measurement is used to determine the relationship between dependent and independent variables. Moreover, it is widely employed to fit a forecasting model. Many techniques have been developed for carrying out regression analysis.

Multiple regression is a statistical approach utilized for modeling with more than one independent variable or feature (Osborne, 2000). The correlation between variables is examined in this model to screen out the effect of each independent variable on the dependent variable or response, and to describe what variables have conceptual senses in a forecasting. Moreover, interaction effect is a common phenomenon in regression analysis, when the interaction of one or more independent variables have a high influence on the performance of the model and forecasting the response value. Hence, in most problems, interaction terms are significant in statistical concepts and modeling (James, Witten, Hastie, & Tibshirani, 2013).

In 1982, Bean proposed a powerful multiple regression model for forecasting student behavior by deploying the interaction term (Bean, 1982). Later on, in 1986 Smouse and colleagues employed an efficient forecasting model with the mentioned method in biology research (Smouse, Long, & Sokal, 1986). It is widely acknowledged that weather forecasting and seasonal climate assessment (Krishnamurti et al., 1999), bike rental demand (Ji, Cherry, Han, & Jordan, 2014), brain research (Klein, Foerster, Hartnegg, & Fischer, 2005), and more scopes are feasible through multiple regression along with the strong forecasting results.

2.2.2 Regression Tree

In recent years, most of the research in statistical learning has been located on non-linear methods. However, non-linear methods often have drawbacks such as low interpretability power in comparison with linear models, but they are adept in discovering pertinent interaction among variables and accurate results (James et al., 2013).

Within most research efforts, decision tree and regression tree have become popular as tree-fitting models (Breiman, Friedman, Olshen, & Stone, 1984). Decision tree is used when the dependent variable is categorical, which is out of the scope of this study. Regression tree can accommodate continuous dependent variables. Since 1991, regression tree has been started up as a powerful model in forecasting problems (Karalic & Cestnik, 1991).

Briefly, tree-based regression is grown by applying greedy algorithms to speedy split that recursively partitions the training set into successively tiny subgroups (Kohavi & Quinlan, 2002). Consecutively, splits are examined for all independent variables, and ultimately the best submission is considered by measuring the impurity by indicating homogeneity in the lower level of the tree.

Augustin and his crew showed that regression tree is able to evaluate risk forecasting in medical scope as an alternative to various models by presenting accurate forecast models (Augustin et al., 2009). The results coming from recent successes show that regression tree, despite its shortcomings, still has special popularity among ecological researchers for exploring interaction and pattern recognition (De'ath & Fabricius, 2000).

2.2.3 Random Forests

Random forests methodology is known as one of the impressive learning models. This method was proposed by Breiman in 1984 which can be used to fit forecasting models for regression and classification problems. An early technique is called bagging which was introduced by Breiman in 1996 and it is a general approach that can be applied in many machine learning methods. Bagging is a method for creating multiple version sets from training set and it helps to improve the model by using each of the sets for training the model (Breiman, 1996). Therefore, bagging contains several trees which most or all deploy the stronger feature in the top split. Breiman has shown another technique in 2001 which only considers a subset of the features for each split. Hence, in all trees the strong feature is not only considered feature, while other features have chance to be in the top split (Breiman, 2001).

Random forests are used in many fields in order to make a robust forecasting model which outperforms artificial neural networks model (Palmer, O'Boyle, Glen, & Mitchell, 2007).

2.2.4 Gradient Boosting,

Gradient boosting is one of the fundamental advances in statistical tools with high adaptability along with elegance and simplicity, which is applicable for both regression and classification for a vast scope of problems and various supervised learning models (Friedman, 2001). This method led to understanding and tuning of model's parameters which can be used to attain the best forecasting model (Kuhn & Johnson, 2013). Due to its numerous benefits, researchers and developers such as Gilberto Titericz, have become a fan of gradient boosting who previously were eager to neural networks. Persson development team drove gradient boosting as an efficient machine learning model in a multi-site framework for fitting a forecasting model of future power generation in Japan (Persson, Bacher, Shiga, & Madsen, 2017). Performance models were evaluated in comparison

with benchmark models; thereby, the results showed that gradient boosting surpassed other models. Robustness and flexibility as criteria were also considered.

2.2.5 LSTM Recurrent Neural Networks

Artificial neural networks (ANNs), as alternative methods of statistical approaches, have drawn substantial consideration in variant scopes, containing computer science, statistics, business, and even diagnosis or treatment. The initial development was inspired by the human brain's neural format, which generate and explore new models in learning (Zhang, Patuwo, & Hu, 1998). ANNs are scattered as one of the major efficient forecasting approaches in several applications. Lapedes and Farber were a pioneer in developing forecasting applications of ANNs (Lapedes & Farber, 1987). In this knowledge, various architectures have been expanded with different neurons arrangement. The long short-term memory based on recurrent neural network is most widely for sequential dataset.

Recurrent Neural Networks are popular networks that have proposed supreme premise in major tasks. In particular, the mentioned model is accommodated with historical, sequential and even time series datasets. RNNs as an extension of neural networks are also based on the premise that the networks handle consecutive information and the connection forms which are in oriented cycles state (Witten et al., 2016). The term of Recurrent in RRNs implies implementing the same role for every member of a sequence, and the extracted result depends on the prior actions. This situation in RNNs proves the memory simply in this model, which retains informative results in each period (Chung, Gulcehre, Cho, & Bengio, 2014). Although, Bengio et al. realized that RNNs are not able to hold informative asset in long-term which impresses negative effect on model performance (Bengio, Simard, & Frasconi, 1994).

Researchers struggled for a while to overcome this weakness and gain an effective execution. Hochreiter and Schmidhuber designed LSTM to enhance the capability of RNNs for learning in long-range dependency (Hochreiter & Schmidhuber, 1997). Lately, Kang and Chen employed LSTM recurrent neural networks for forecasting traffic streams, and yielded significant achievements (Kang, Lv, & Chen, 2017). Recognizing a temporal pattern of traffic is indispensable for the transportation system, because undoubtedly the obtained forecasting models and results could inspire urban development.

2.2.6 GRUs Recurrent Neural Networks

GRUs as a new version of LSTM is recently provided by Cho et al. (Cho, Van Merriënboer, Bahdanau, & Bengio, 2014). GRUs address problems similar to LSTM with low memory demand. It technically means that there is no necessity to employ detached memory cells in process of capturing information inside units. Therefore, considering low-cells in described model led up to more efficiency than LSTM. Note that GRU is plainer and more computational than LSTM (Chung et al., 2014). In 2017, Chen et al. demonstrated that recurrent neural networks with GRU blocks, can be applied in order to fit a model to forecast bike demands with concern about distribution of bikes in sharing service. The proposed model in this study was compatible with time series dataset (P.-C. Chen, Hsieh, Sigalingging, Chen, & Leu, 2017).

2.3 Performance Metrics

Fitting a model for forecasting future data is one of the steps of forecasting process. The next step is measuring the quality of the model. In order to assess the desirability of machine learning algorithms on a determined dataset, performance metrics assist to measure models' quality to fit in accordance with observed data. Furthermore, these metrics are highly substantial for evaluating and comparing different algorithms to each other based on various criteria such as stability, robustness to noise and so more, which depend on the type of problem. For instance, in regression models, mean squared error (MSE) is usually used, given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.1}$$

where y_i is referred to actual value for given data, and $\hat{y}_i$ is the value returned by the model or predicted value. In equation (2.1), training data is used for generating the model, and test data is applied to examine the model with-which difference is adverted to the accuracy of predicted model.

Due to the model variety, decision for selecting the best metric for the model assessment is usually not easy. Some of the other statistical tools for measuring the goodness of the result are as following: root mean squared error (RMSE), adjusted R^2, Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallow's C_p, and more. Each of them is adjusted to judge specified methods and issues (James et al., 2013)

CHAPTER 3 METHODOLOGY

The proposed methodology is explained in this chapter. Section 3.1 discusses about the goal of this thesis. Section 3.2 introduces the types of dataset, consisting of Communauto, holidays and weather dataset in CSV format. This section also covers selection of the most relevant variables with the considered responses, individually for each dataset. Thereafter, section 3.3 represents the foremost practice step to visualize given data in order to clean and prepare the raw data for analyzing, by using Python language programming. Data splitting and its effect on performance of a model are described in section 3.4. Section 3.5 is about adopting some machine learning techniques which are used in this project. The learning procedure is explained by how candidate models can be practically applied to the dataset to uncover patterns. Moreover, in section 3.6, performance of the techniques is evaluated through test or validation set to forecast desired response (consumed hours and mileage). The summary of the processes of this study is shown in Figure 3.1.

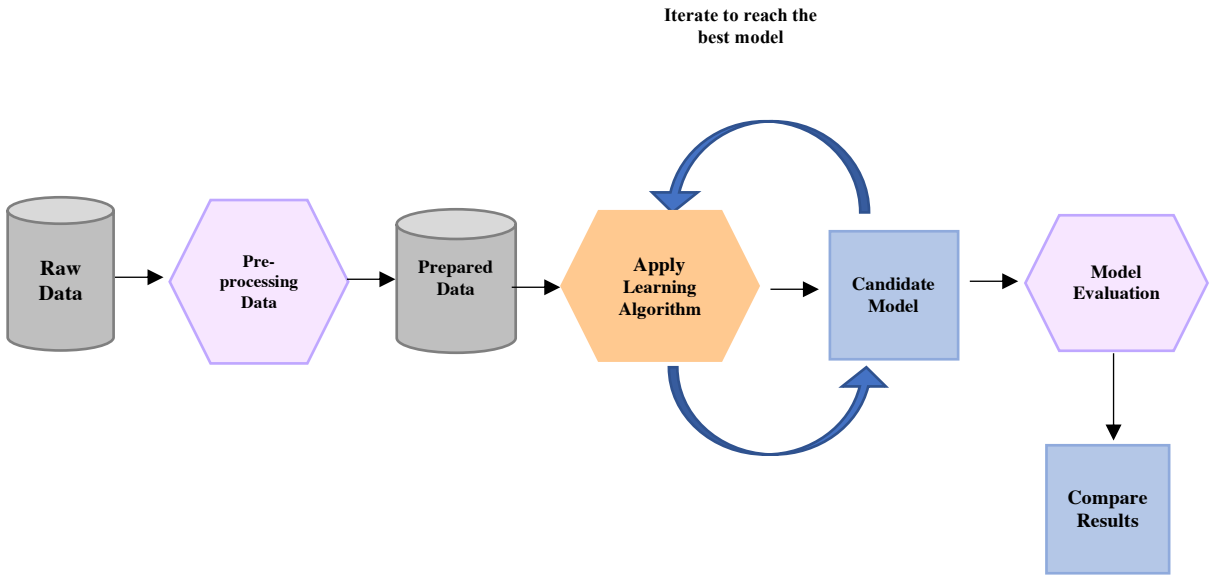

Figure 3-1: The design summary for forecasting process

3.1 Problem Review

The main goal of this thesis is to build some forecasting models by supervised machine learning techniques, based on the historical dataset. This project is carried out as a comparative analysis for forecasting algorithms. The principal concern of this study is forecasting the consumed hours and mileage as desired responses or outputs of this study regarding to different factors. Initially, the model will be built without involving additive factors, thereafter, they will be engaged to investigate whether they improve the performance of the models or not.

Principally, in order to have the best model with lowest error rate, some operations should be employed to prepare the data, take the informative variables, and build a model.

3.2 Data Description

In order to interrogate the factors affecting consumed hours and mileage in the Communauto carsharing system as the targets of this research, three different sources of data are applied to serve. Applied data includes: (1) Communauto carsharing services, (2) historical weather conditions in Montreal and (3) holiday information of Quebec Province. The following sections are provided to explain every sample in detail with different features and several types of values. Moreover, data preparation procedures such as feature dropping or variables augmentation are presented in each data description. It is noteworthy that the pertinent datasets to regular and free-floating services are subcategorized under Communauto dataset section.

3.2.1 Communauto Dataset

These datasets have been provided by Communauto carsharing network based in Montreal. As mentioned in the first chapter, this company has been providing regular or station-based, and free-floating (called “Auto-Mobile”) services. The dataset is gathered by both services that can be distinguished through the column assigned to service. It means that accumulated rows by AUM in the column labeled as service refer to the free-floating “Auto-Mobile” service, and the other part which is displayed by REG is related to the station-based or regular service.

Both services have the same variables as the other one and the only difference is the variable which is dedicated to the number of active stations. There is no station in free-floating service because the users take the fleet of scattered vehicles which are available in the authorized zone.

The given dataset brings CSV (comma-separated values) file of carsharing trip histories, which consists of 2982 observations with 10 variables. The recorded information in Communauto dataset contains the number of stations, vehicles, users, reserved and free cars, and total consumed hours and mileage on each day which are conducted from 2010 to 2016 for regular service and 2013 to 2016 for free-floating service.

Each variable is typically called "feature" in this thesis. As previously mentioned, the main goal of this project is forecasting the consumed hours and mileage in the company of interest. Therefore, the variable labeled by *nbHeuresVehYMD* (consumed hours) and *distTotaleKmReservation* (mileage) are considered as individual dependent variables or output which is referred as response in this report.

It should be noted that each response as a target of this research is uniquely placed under the study. Hence, the regular-service or REG dataset and consumed hours, as response, are considered for the forecasting process. The approach is the same for mileage as another target and Auto-Mobile service dataset in this research. The results are demonstrated in this report.

- **Feature Importance**

The Communauto dataset is decomposed into two particular sets: 1) Auto-Mobile (AUM) 2) regular (REG). The first 1005 rows include AUM observations from 2013 to 2016, and the rest include 1977 rows related to REG information between 2010 and 2016. Table 3.1 illustrates the details of this dataset.

Table 3-1: Description of Communauto dataset

Feature Name	**Type of Data**	**Description**
YM	Date	Date which includes month and year
YMD	Date	Date which includes day, month and year
nbEmpruntDebutYMD	Numeric	Number of reserved cars on specific day
nbEmpruntFinYMD	Numeric	Number of free cars in end of the day
nbHeuresVehYMD	Numeric	Consumed hours on specific day
nbVehiculesActifsYM	Numeric	Number of active vehicles in specific month
nbStationsActivesYM	Numeric	Number of active stations in specific month
nbUsagersActifsYM	Numeric	Number of active users in specific month
distTotaleKmReservation	Numeric	Consumed mileage on specific day

The principal task is to remove the feature that do not have an influence on this study. To this end, in both of the grouped data, the first feature labeled as *YM* is eliminated, because the next labeled as *YMD* represents more detailed aspects of the timestamp information to identify the year, month and day. Due to the *YMD* feature, observations are the sequence of historical spanning data, since this research is aimed to forecast the desired responses in different month and daytime scale. Therefore, *YMD* feature in a couple of processing can be isolated into the month and week of day variables. The column including the number of months is determined by values between 0 to 12 in the specified year, where 0 refers to the month of January. The week of day variabel is filled with day types, which are represented by a number between 0 to 6 in weekday observations, in-which Monday is considered as the start day (day 0). The recorded date can assist to answer the following questions: Whether the amount of consumed hours and mileage will be depended on different months, and which months will be more aggregated? What kind of days in this service has the most frequent usage, work day or weekend?

Moreover, none of the variables labeled as *nbEmpruntDebutYMD* and *nbEmpruntFinYMD*, referring to number of reserved car and free vehicles at the end of the day, contain effective information about the desired responses under this study. The irrelevant features may significantly degrade the performance of a forecasting model. In such a case, these features can be discarded from the initial dataset without performance deterioration in forecasting process.

The available data related to AUM or Auto-Mobile service has the same features except the variable labeled as *nbStationsActivesYM*. Notice that all the operation lines above, are applied to preparing the AUM set.

After dropping those features, basically the dataset contains six main features with nature of date, character and number: *YMD*, *Month*, *WeekDay*, *nbVehiculesActifsYM*, *nbStationsActivesYM*, and *nbUsagersActifsYM*. Table 3.2 describes comprehensive information of features and response variables in Communauto dataset.

Table 3-2: Description of the most relevant variables of Communauto dataset

Feature Name	**Type of Data**	**Description**
YMD	Numeric	Year
Month	Numeric	Month
WeekDay	Numeric	The type of day in each week
nbVehiculesActifsYM	Numeric	Number of active vehicles in specific month
nbStationsActivesYM	Numeric	Number of active stations in specific month
nbUsagersActifsYM	Numeric	Number of active users in specific month
nbHeuresVehYMD	Numeric	Consumed hours on specific day
distTotaleKmReservation	Numeric	Consumed mileage on specific day

The considered features interfere with the process of creating forecasting models in order to forecast the consumed hours and mileage in the specified time as the targets of this study.

3.2.2 Weather Dataset

The historical weather dataset represents the overall conditions in Montreal that is recorded in CSV format during the period from January 1, 2008 to December 30, 2017. Each observation is incorporated with daily weather. This dataset is composed of several features, such as maximum, minimum and mean of temperature, the amount of rain and snow during the day and more.

- **Feature Importance**

The columns pertinent to rain and snow must be ignored, because most of them are filled by null values and do not provide much information. Therefore, to avoid imposing the complexity in analyzing and modeling procedure, the average temperature is considered as a part of data under this study. The varibale labeled *WeatherMeannTemp* is referred to the average temperature in dataset. The recorded observations in this feature are numeric, which are restricted within the range

between -25 to 30 in Celsius scale. In this case, the values in the column *WeatherMeannTemp* can be converted to two levels of possible dummy variables, zero and one.

3.2.3 Holiday Dataset

In Canada, holiday depends on the province. Typically, each province has its own holiday periods. The major public holiday information of Quebec province is presented in CSV file between January 1, 2010 to February 31, 2016. Saturday and Sunday are counted as holiday in obtained dataset based on a hypothesis that consumed hours and mileage are different on workday and weekend. Hence, the data model just distinguished between holiday and non-holiday or working days. In two categorized classes, holiday times are shown by Yes and the rest are represented by No.

3.3 Data Preprocessing

Usually, raw data is seldomly suitable for analysis and feeding algorithms. Such data have significant effect on the performance of a model. The preliminary and critical task is to investigate the quality of the given data. Most often, real data is congregated by incomplete, inconvenient, unreliable and distorted information. In order to do this, it is required to clean up the initial data from redundant and irrelevant information. Data preprocessing knowledge covers several tasks with the aim to convert raw data into an applicable format and prepare the data before formal analysis. In fact, preprocessing knowledge trace potentially significant information in initial data, understand the complexities and discover good ways to handle such issues. Generally, data preprocessing would fairly lead to achieving significant performance of supervised models (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

3.3.1 Null Values

Missing or unknown values are an unavoidable problem in raw data and can have a substantial impact on final results. There are varieties of sources to implicate the data into issues such as mistakes in measurements, misrecording of data and so on (Kantardzic, 2011). Available sample received on behalf of Communauto company contains some of the missing observations which are distinguished by 0 and NA. To resolve this issue, because the missing data is restricted to a few cases and the quantity of data is enough for analysis and query, the rows which meet this state are dropped from the dataset (Scheffer, 2002).

3.3.2 Detecting Outlier: Box Plot Diagram

Outlier is as an unusual observation that is stated outside the overall pattern of data model and it is drastically inconsistent with other observations. Such outliers typically have an unpleasant influence on the results made from the study. This issue can arise due to a variety of either faulty recording or empirical error. In other words, outliers are awarded to the observations that the response y_i is weird given the feature x_i (Kuhn & Johnson, 2013). These types of unusual values require precise analysis before deciding whether they should be eliminated from the population under study, or not.

Box plot diagram is a convenient method in order to identify outliers and eliminate them from data, where appropriate (Zani, Riani, & Corbellini, 1998). This plot is presented by quartile and interquartile or IQR for determining the lower (Q_1) and upper (Q_3) quantiles, and also lower and upper limits. To this end, if the observations in the dataset lie within the lower and upper limits, they are considered as normal observations. Otherwise, they are called outliers. Thus, they are not eligible to be used for further study. The IQR range is the extension of the middle 50 percentage of data values.

$$\text{IQR} = Q_3 - Q_1 \tag{3.1}$$

$$\text{Lower Limit} = Q_1 - 1.5\ \text{IQR} \tag{3.2}$$

$$\text{Upper Limi} = Q_3 + 1.5\ \text{IQR} \tag{3.3}$$

3.3.3 Encode Categorical Variables

The weather and holiday datasets have categorical or symbolic variables, in-which observations capture non-numerical labels rather than numerical. Some algorithms and Python libraries are not able to work with categorical data directly. These observations need to be converted into numerical nature. For example, regression analysis and Python library 'sklearn' require feature in number. In order to overcome this problem, dummy coding can be applied for constructing a categorical variable into a numerical variable that takes one of two conceivable dummy numerical values. For example, based on the feature observations, a new one takes one or zero such as the following form (3.4).

$$x_{ij} = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{3.4}$$

In general, x_{ij} represents the value of the jth variable for the ith observation.

3.3.4 Normalized Data

In LSTM and GRU based on RNNs, algorithms are known to provide reckless forecasts on the observations with different format and scale (Russell & Norvig, 2016). Therefore, normalized the features must be carried out to lie in a fixed range. This range is usually from zero to one based on maximum and minimum observed values in each column. As shown in the following equation, the observed value (x) is subtracted by minimum value, and the result is divided by a range between the maximum and minimum values.

$$\mathcal{Z} = \frac{x - min\ (x)}{max\ (x) - min\ (x)} \tag{3.5}$$

3.3.5 Data Visualization

Graphical representation of data is called data visualization. It is a crucial step in forecasting problems. Generally, the data is displayed by visual tools such as plots, graphs, charts and more (James et al., 2013). Visualizing data dedicates valuable insights about data and further exposes to see visual patterns of the dependency of every pair of variables. Pearson correlation is most widely used in order to overview data. In this end, linear association between two variables (X and Y) by the following equation (3.6) is measured, where $\bar{x}$ is the mean of X variable and $\bar{y}$ is the mean of Y variable.

The result is a degree of dependency between variables. The coefficient correlation is in a range of value $+1$ to -1. Value of zero indicates that there is no relationship between variables, and it can be ignored from the data. A value close to +1 or -1 indicates strong correlation between those variables.

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.6}$$

3.4 Data Splitting: Forward-Chaining

A common method to avoid overfitting is to split data into non-overlapping separated subsets, which are called train and test sets. This function assists to train and evaluate forecasting models. Initially, an algorithm is learned to forecast based on previous observations, which are known as training data. Training data employ to build forecasting models. Generally, in creating a model, we do not trace how well the trained algorithm works on a training set. Rather, the efficiency of a model on new observations or test data is interested, because most of the time, the model works much better on training data than on test data, which leads to the phenomenon of overfitting the data (James et al., 2013). In order to prevent from getting stuck at such a problem, the model will be evaluated based on the test set.

In this project, observations dependent on time. Therefore, it is not possible to keep a model blind about this dependency. Due to this kind of data in this project, random separation does not work. The data should be cut off based on time. For example, older dates are determined as training set and recent dates are considered as test set.

A common approach is used to produce a better estimation of error forecast of a model, which is called forward chaining and referred as rolling-origin. In this technique, the full data is split into k-fold train/test subsets (Bergmeir & Benitez, 2011). Each time, a specific year is considered as a test data and the previous year or years are taken as a training set. This process keeps moving forward through the data, until it covers all the dataset. The trained models with each test data produce new predicted values. Therefore, the difference between predicted and actual values is computed by the considered metric performance in this study. Afterwards, the final error is obtained by taking the average of k folds results.

3.5 Data Modeling

The main goal of this study is to compare sets of supervised machine learning algorithms, such as multiple regression, regression tree, random forests, gradient boosting, and LSTM and GRUs recurrent neural networks. They are reputed as the most accurate forecasting models. Therefore, these algorithms will be implemented to create the models. As discussed before, the target of this project is applying the mentioned supervised learning techniques to forecast the amount of consumed hours and mileage, individually, in Communauto carsharing operator in time scale.

3.5.1 Multiple Regression

One of the extensions of simple linear regression is called multiple regression. In particular, it is an effective tool for forecasting problems with estimating the relationships between independent variables or features and quantitative dependent values or output. Multiple regression does not assume a linear relationship between response and features.

General format of this model is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (3.7)$$

Where if there are p features in dataset, X_p refers to the pth feature, Y is the response value and ε is a random error term. β is regression coefficient which is unknown. This method is traced to find the best estimation that the model fits the available data well and minimizes the error rate.

It is often more convenient to use matrix notation:

$$Y = X\beta + \varepsilon \quad (3.8)$$

Y is $n \times 1$ vector:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

X is $n \times (p+1)$ matrix:

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1\mathrm{p}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{\mathrm{n}1} & \cdots & x_{\mathrm{np}} \end{bmatrix}$$

β is $(p+1) \times 1$ vector:

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{\mathrm{p}} \end{pmatrix}$$

ε is $n \times 1$ vector:

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The estimated unknown coefficient is achieved by minimizing the sum of square errors method. This is implemented by taking the first derivative of sum of the squared errors with respect to estimated coefficient or $\hat{\beta}$. Then, set the derivative is set to equal zero (Lang, 2013).

The sum of the squared errors:

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= (y - X\hat{\beta})^T (y - X\hat{\beta})$$

$$= y^T y - X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T \hat{\beta}$$

Take derivative with respect to $\hat{\beta}$:

$$\frac{\partial (y^T y - \mathrm{X}\hat{\beta} - \hat{\beta}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{y} + \hat{\beta}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}} X\hat{\beta})}{\partial \hat{\beta}} = 0$$

$$-2\mathrm{X}^{\mathrm{T}} y + 2\mathrm{X}^{\mathrm{T}}\mathrm{X}\hat{\beta} = 0$$

$$\mathrm{X}^{\mathrm{T}} y = \mathrm{X}^{\mathrm{T}}\mathrm{X}\hat{\beta}\, y$$

Therefore:

$$\hat{\beta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1} X^T y \tag{3.9}$$

Regression coefficient refers to the change in Y associated with a one-unit change in the respective independent variable. Statistical tests try to find whether each coefficient is significantly different from zero.

Furthermore, linear correlation coefficient is used to determine the dependency of variables, which is in range of $+1$ or -1. When the value is close to zero, it means there is no linear relationship. If correlation goes to near minus or plus one, it indicates a strong linear relationship between those variables. Multiple regression examines the subset or full features to come up to the combination of them, which leads to the best result. In some situations, regression analysis cannot detect a relationship, because of circle form. For this reason, creating plot is as an alternative way to

mapping a causal effect relationship between features and dependent variables. Therefore, in some circumstances, more complex model may contain variables with higher powers. For instance:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^i + \cdots + \varepsilon_j \qquad i = 2,\ldots \tag{3.10}$$

Sometimes, two or more variables have a significant relationship. Therefore, the effect of their interaction is considered in procedure of creating model. For example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12}(X_1 X_2) + \cdots + \varepsilon_j \tag{3.11}$$

Therefore, the combination of these terms creates multiple regression. The associated p-value and individual t-statistics is applied to test whether a variable contributes significantly to the fitted model. The level of significant p-value is considered 0.05 for inclusion of a variable in the structure of a model and rejection the null hypotheses (H_0*):*

$$H_0 : \beta_j = 0$$

The results of regression are easily interpretable, and the pattern can be recognized by plots (James et al., 2013). But, this model is sensitive to outliers, and such points can seriously affect the final results.

3.5.2 Regression Tree

Regression tree is first proposed by Breiman and colleagues in 1980 (Breiman, 2001). Regression tree is an alternative approach of nonlinear regression, and it is most widely used in regression problems. It means that this model can be constructed as a forecasting model in dataset, where dependent variables or output lying in a numeric range. The implementation of forecasting action is set by partitioning the data into smaller regions. The obtained segmentations are more manageable and can be summarized in a tree with terminal nodes or leaves and branches. Then, the operation of partitioning splits the subdivisions again, which is called recursive partitioning. This tree allows variables to be a mixture of categorical, continuous, sparse, skew, etc. Even this model is compatible with missing values in the dataset. Structure of the tree is able to adapt to large data with no requirement to know the correlation between each individual feature variable and response (Strobl, Malley, & Tutz, 2009).

Figure 3.2 shows a top-down tree, which is started from root node with querying a sequence of questions about features and places a feature as a representative at this node. Decision node refers

to sub-node, when it is divided to further sub-nodes. In contrast, when a node does not divide, it is called leaf or terminal node. The nodes are accumulated with questions, and branches between nodes are labeled by the answers. Next question depends on answers to overhead questions. Finally, each part of the observations is assigned to subset with similar response values, and the rest are assigned in other subgroups. So, the goal of all the efforts is to achieve an optimal partitioning of the data.

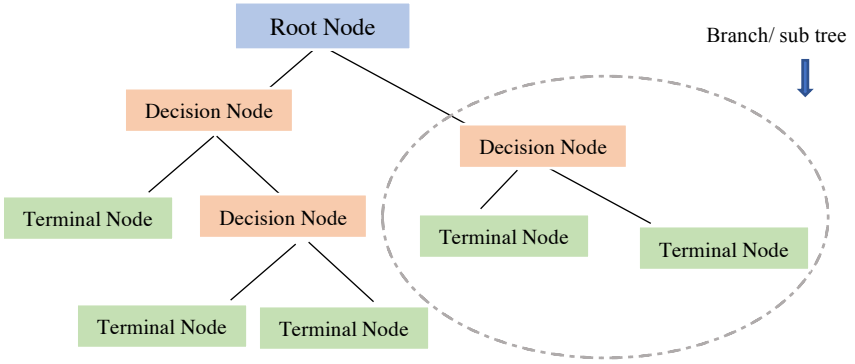

Figure 3-2: A recursive partitioning tree

There are different splitting criteria in regression tree, in-which MSE (mean squared error) as one of them is popular. Firstly, in order to perform division, one feature as root node which leads to the lowest MSE, is selected. In each node, $\hat{y}_i$, as the response mean of the training sample, and MSE are calculated separately within node (depicted in (3.12)). For example, observation $X = x$ when $x \epsilon N_j$, $\hat{y}_i$ is the predicted value, N_j refers to jth node.

$$\hat{y}_i = \frac{1}{c} \sum_{i=1}^{c} y_i \tag{3.12}$$

Thereafter, the value with significant MSE will be chosen to best split the tree at that particular step. The process will elaborate for other observations in training set.

Ideally, the goal of this method is finding a partition with lowest mean squared error for given regression problem. Recursive splits on dataset continue upon reaching the state of defined

maximum depth or another stopping parameter. Consequently, the model is seeking a tree that minimizes the MSE given by:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{i\epsilon N_j}(y_i - \hat{y}_{N_j})^2 \quad (3.13)$$

This process may have a good result on training set and poor performance on test set. Usually, tree result is too complex, and that is why it cannot behave well on test set. A tree with fewer split is more interpretable with lower variance. Therefore, pruning the grown tree, T_0, is as a strategy to obtain a subtree with fewer nodes. In order to obtain a subtree with the best result, we can specify different maximum numbers or depth of levels below the root node and evaluate the results by the criteria metrics.

This model is really sensitive to outlier, and smallest change in data can modify the overall structure of the optimal tree.

3.5.3 Random Forests

Random forests model, as one of the tree-based models and well-known machine learning techniques, is proposed by Breiman (Breiman, 2001). This model is in contrast to regression tree, which uses full features or variables at each step of split. It applies the random selection of features as candidates and chooses the best split among those variables (Joelsson, Benediktsson, & Sveinsson, 2005). In other words, just a few of the available features are used at each split in the tree. MSE is considered for regression problem as an indicator of the best split and evaluating model. Considering that one of the variables is more important than the others, random forests do not employ this variable as strong feature. It also allows other features to have more chance. Typically, random forests recommend using $m = \log(p)$ or $m = \sqrt{p}$ instead of all features, where p refers to total number of variables in a dataset. Along with the stability, highly performance and efficient characteristics, it is compatible with a high-dimensional data where it includes a large number of correlated variables and non-linear problems. Random forests are also able to deal with a large number of missing values and estimate them by maintaining high performance. However, random forests are fast to train, but the procedure of making forecasting on extensive problems is drastically slow. Furthermore, random forests produce quite flexible models, but it is somewhat less interpretable. Therefore, it is basically suitable when only high performance is interested.

3.5.4 Gradient Boosting

In 2002, Friedman developed an adaptive method for many types of applications, which is called gradient boosting (Friedman, 2002). Gradient boosting as an ensemble of tree-based models, and a machine learning technique to solve regression and classification problems. It is widely used and highly effective and can be applied to many machine learning methods. In this model such as the other learning techniques, a loss function is defined to optimize it. Loss function represents the difference between predicted value and real value, or error residuals. Gradient boosting consists of a series of additive models which distinguishes strong learners from weak learners. Iteratively, it adds each model and computes the loss function. Gradient boosting is intended to update the model in order to minimize the error. In the first iteration, full samples are used to fit a simple model. Then subsample is selected randomly, then it is applied instead of completing data to fit a weak learner and calculate loss function. Each new model or tree helps to improve error made by formerly trained tree. This process is iteratively applied to make a weak learner and add to strong learners until the residual of error gets stabilized. Generally, three parameters involve in gradient boosting method: depth of trees, number of trees and learning rate or shrinkage. Number of trees is like maximum depth in regression tree, which assist to avoid overfitting problem. The learning rate is a small positive number that reduces the process and allows more different trees to correct the residuals.

The model may get stuck into overfitting problem, when number of trees is too large. In general, gradient boosting learns slowly, which tends to perform well.

3.5.5 Recurrent Neural Networks

Recurrent neural networks (RNNs) with directed cycles form are a powerful and robust extension of neural networks. RNNs have internal memory and they are able to keep internal state, which helps to produce a precise forecasting model. This network uses the output of each neuron as an input for the same neuron at the same step (Witten et al., 2016). In other words, it has two inputs, the recent past and the present, to feed to the next element of the sequence. This contains helpful information about what is happening next. Bengio et al discovered that RNNs are not suitable for capturing long-term dependencies(Bengio et al., 1994). Therefore, researchers have tried to find a way for resolving this issue.

3.5.5.2 Long-Short Term Memory

One of the effective approaches to overcome long-term dependency problem is long-short term memory (LSTM) with memory blocks for layers in RNNs. Initially in 1997, this model was developed by Hochreiter and Schmidhuber (Hochreiter & Schmidhuber, 1997). RNNs employ LSTM blocks to improve their performance when the network requires to learn long-term connections. It assists to store specific feature of input for long term. In this state, the network may face vanishing gradient issue, and LSTM can reduce this difficulty (Hochreiter & Schmidhuber, 1997). Vanishing gradient causes to keep weights without correction for long periods of time and stope learning in layers. Figure 3.3 depicts the complex architecture of LSTM unit, in-which at each step three different gates exist: input gate (i_t), output gate (o_t) and forget gate (f_t).

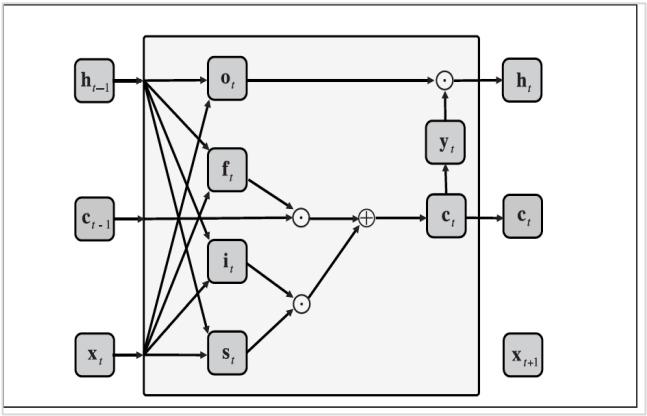

Figure 3-3: Architecture of Long Short-Term Memory unit (Witten et al., 2016)

At time *t*, the network takes three inputs. X_t is the input vector of current block, hidden state h_{t-1} is the output from previous block, and c_{t-1} is the memory from the previous block which is the most important input. Furthermore, h_t is the output of the current block and c_t is the memory of the current block. Therefore, the block at time *t*, makes decision by considering current input, previous output and memory; then it creates a new output with its memory. The activation function is sigmoid in this neural network and it is one of the activation functions in neural network techniques which squishes values between 0 and 1. Figure 3.4 shows the form of this function. X_t by their own weight matrix, and at time *t-1*, hidden state h_{t-1} by its own U matrix, and their bias vector are passed to the sigmoidal non-linearity layer to create i_t and o_t as an input and output (equations (3.14), (3.15) and (3.16)).

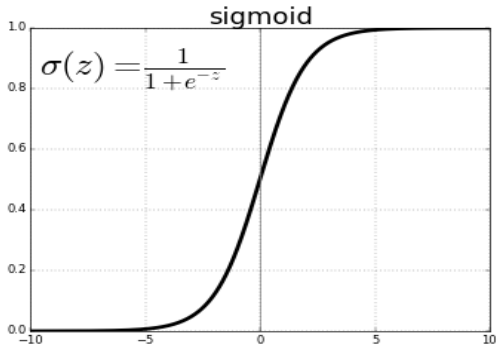

Figure 3-4: Sigmoid activation function (Abhijit Mondal, 2017)

Considering the product of sigmoid, 0 means not important and 1 is important, and it goes through the network.

$$i_t = sigmoid\ (W_i x_t + U_i h_{t-1} + b_i) \tag{3.14}$$

$$o_t = sigmoid\ (W_o x_t + U_o h_{t-1} + b_o) \tag{3.15}$$

$$f_t = sigmoid\ (W_f x_t + U_f h_{t-1} + b_f) \tag{3.16}$$

Then the results of input gate pass the tanh layer which produces new candidate vectors as potential input, s_t, given by equation (3.17). The tanh activation function in the following equation (3.18), squishes values between -1 and 1. Then, the result of input gate is used to identify whether potential input is adequately important to store into the memory unit, c_t, shown in equation (3.19).

$$s_t = tanh\ (W_i x_t + U_i h_{t-1} + b_i) \tag{3.17}$$

$$h(x) = \frac{exp(x) - exp\ (-x)}{exp(x) + exp\ (-x)} \tag{3.18}$$

$$c_t = f_t \times c_{t-1} + i_t \times s_t \tag{3.19}$$

Then forget units f_t prevent inappropriate information accumulation in memory cell, and it allows to content to be erased. The final phase, output gate decides whether y_t (the stored content in memory units, which are transformed by activation function) should flow to hidden units h_t or not. The hidden layer at time *t* is gained by the following equation (3.20).

$$h_t = o_t \times \tanh(c_t) \tag{3.20}$$

3.5.5.2 Gated Recurrent Units

GRUs similar to LSTM unit, they are a specific extension of RNNs. They adjust the information inside the unit, but unlike LSTM do not contain separate memory cells. Another difference with LSTM is in memory capacity, which is appropriate for problems with the requirement of the lower memory or smaller dataset (Chung et al., 2014). This model is proposed by Cho et al (Cho et al., 2014). Figure 3.5 indicates the architecture of GRUs along with gates.

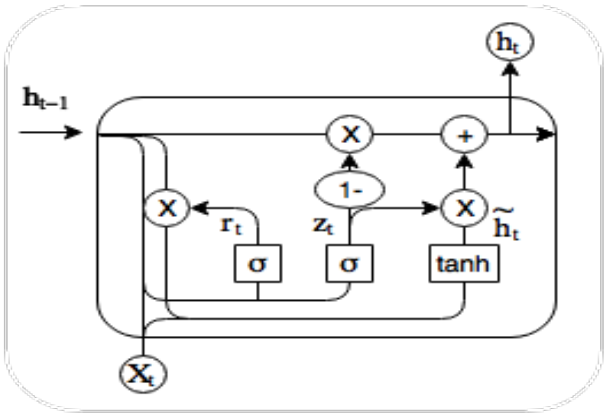

Figure 3-5: Architecture of Gated Recurrent Units (Danny Mathew, 2018)

In this model, the network is consisting of two types gate: rest gate (r_t) and update gate (z_t), which makes this model more efficient than LSTM. Update gate (z_t) has the same task as the forget gate and input gate of LSTM. This gate decides how much of previous memory to keep around. The forget gate determines the vector of stored information or contents. The reset gate defines how to combine new input with previous value. The equations (3.21) and (3.22) refer to update and rest gates.

$$z_t = sigmoid\,(W_z x_t + U_z h_{t-1} + b_z) \tag{3.21}$$

$$r_t = sigmoid\,(W_r x_t + U_r h_{t-1} + b_r) \tag{3.22}$$

There is no input gate in GRU. Therefore, the entries of previous hidden state h_{t-1} and input values X_t are considered as input. The rest gate stores the prior memory in the hidden state and new input as a vector and determines how they combine together. Computation of the rest gate is similar to update gate with sigmoid function. The hidden state $\widetilde{h_t}$ is a new memory at time t similar to potential input in LSTM (3.23). In this case, previous hidden states are computed by the equation (3.24).

$$\widetilde{h_t} = tanh\,(Wx_t + U_i(r_t \cdot h_{t-1} + b_i)) \tag{3.23}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t h_t \tag{3.24}$$

3.6 Evaluation Methods

Evaluating the applied algorithms can be done by several performance metrics. This step is an essential aspect of the research, because without evaluating the performance of created models, there is no clue for understanding that the research is done correctly or not. The first principle is, using test set for evaluating the performance of a model rather training set.

There are various alternative measurement methods in order to evaluate the forecasting models. Most commonly, the principal method in regression problems for performance measurement is MSE, and sometimes the root mean squared error (RMSE) is used (James et al., 2013). Some algorithms are compatible with these methods and they behave well. The results of these methods are reported in summary section by default, but the others need to be computed separately. For instance, in regression tree algorithm, one of the default performance metrics is MSE. In order to obtain error measurement RMSE, square root should be taken from the quantity results. The equations of RMSE and MSE are as follow:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.19}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.20}$$

Where y_i is an actual response value and $\hat{y}_i$ is the forecasting value, for the ith observation. When the results of RMSE or MSE are very small, it means that the predicted values are close to real response values. In other words, the model is significantly fitted to the given dataset.

CHAPTER 4 IMPLEMENTATION AND RESULTS

This section presents the implementation of the models, and their results. The results of the process of data cleaning from null values and noisy points are described in section 4.1. Afterwards, in section 4.2 the prepared data are used to train the numbers of supervised machine learning techniques such as multiple regression, regression tree, random forests, gradient boosting, LSTM and GRUs recurrent neural networks. Thereafter, the models are examined by the test set and the results are explained in 4.3 as evaluation part.

As mentioned before, the data related to station-based or regular service is used as Communauto dataset for creating the forecasting model with respect to consumed hours, as desired response variable in this study.

4.1 Prepared Data

In section 3.1, the nature of different datasets and the most relevant features were explored. Some features were dropped in Communauto and historical weather conditions dataset, but instead, new columns were augmented in order to improve the data. The column labeled as *YMD* was decomposed to year, month and week day, and two irrelevant features were removed. Historical weather conditions data under this study contains a feature that indicates the daily average temperature. Furthermore, holiday dataset includes only one column for presenting non-holiday and holiday, where Saturday and Sunday count as holiday times.

As discussed in the chapter 3, the next step after preparing each dataset is to get rid of the issues identified in the dataset, in order to construct them for analyzing and modeling steps. Thus, the following steps are considered as the foremost section, in order to enhance the quality of instance data:

1. Removing null values by removing instances
2. Cleaning data from outliers by using BOX plot diagram
3. Converting categorical data to dummy variables
4. Normalizing data to a value between zero and one for specific algorithms
5. Visualizing data by plotting distribution and correlation
6. Splitting data into training set and test set by forward-chaining

The results deduced from preprocessing the data are explained in each step in the following sections.

4.1.1 Remove Null Values

After the query in REG dataset that represents station based or regular service of Communauto operator, I found that the missed values are related only to the features that represent consumed hours and mileage. Therefore, in this study, it is hypothesized that the rows associated with hour and mileage filled with either 0 or NA, are considered as missing values. Hence, the rows which meet this state, drop from the dataset. In total, there are 13 missed values related to REG data, that are eliminated. As discussed before, most columns in the initial weather data are incomplete, and the best feature is the column that indicates average temperature. The period between 2010 and 2016 corresponds to REG data captured as the final weather data. Since holiday dataset is absolutely free of null values, only the same time period matches both data that are drawn from the given data.

4.1.2 Discard Outliers by BOX Plot Diagram

Both holiday and weather datasets are free of any noisy data or outlier after preparing step. In order to find the outliers in REG data, a column labeled as compare is appended to the prepared data. The additional column is obtained by dividing mileage values over hour observations. The received values via this operation express kilometer per hour ($^{km}/_{hr}$).

Figure 4.1 displays a Box plot diagram of the column labeled as compared to discover the outlier points where the data is suffered from them. Considering the following plot, in total, the points where are gone beyond the lower limit and upper limit as defined range in this method, are 101 observations. This number is 5 percent regarding the total data.

More than 65% of these observations are related to the first day and last day of each month. This problem might be related to the application and Communauto online system. They are as the main source of the collected data in this company, which get into trouble at these times. The reason might turn out to compute users' usage and making an invoice for them at those times. Therefore, the rows related to the identified outliers must be removed from the Communauto dataset. At the end of this section, the column labeled compare is removed.

Besides this, values of the last sixteen rows of the dataset suspiciously have diminished and are not logical. For example, the number of active stations or vehicles is much lower compared to the other months and years. It is assumed to be fictitious information and they are removed as a noisy point from the data.

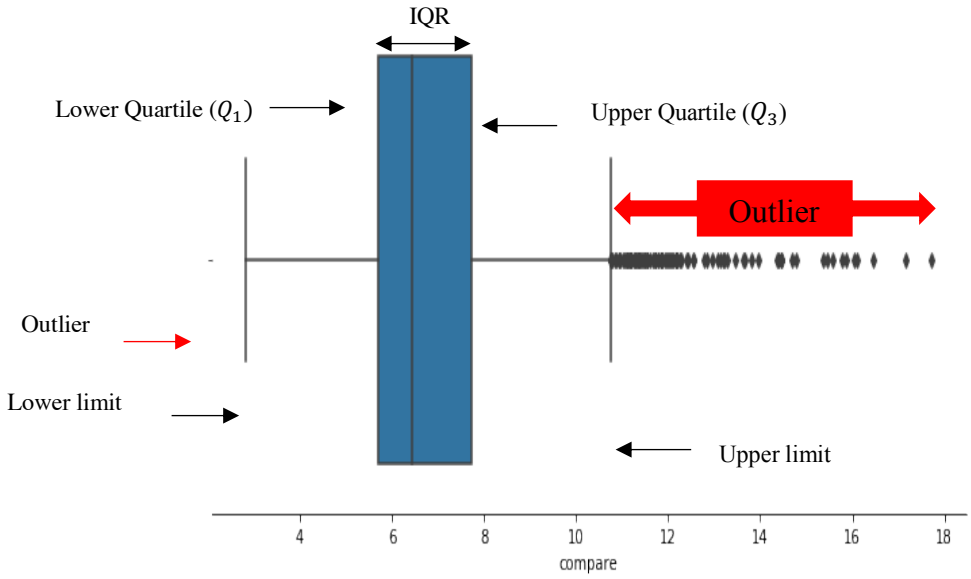

Figure 4-1: Box plot to illustrate the distribution of a kilometer per hour. The lower and upper quartiles and lower and upper limits from the obtained feature (compare) are marked. Moreover, the outliers are shown by red arrow.

The descriptive statistics of continuous features in REG data are shown in Table 4.1. A list of information associated with the provided variables in the following table contains the total number of observations, the mean and standard deviation, the minimum and maximum values and quartile range of each continuous variable, denoted by 25%, 50%, and 75%. As it is given in the table at the first row labeled as count, all the observations of each column have the same number. It means that there is no missing value in the cleaned data.

Moreover, considering the maximum values and the third quartile of each variable, i.e. the maximum value of used hours of vehicles, is 992; while 3rd quantile is 937. It is clearly perceived that there is not much difference between them. Therefore, it means that most of the noisy points

which were identifiable, have been removed from the dataset and the data are now ready for analysis.

Table 4-1: Descriptive statistics of numeric variables in REG (regular service) dataset

	NbHeuresVehYMD	NbVehicules ActifYM	nbStations ActivesYM	nbUsagers ActifsYM	distTotaleKm Reservation
count	1847	1847	1847	1847	1847
mean	907.26	306.55	9475.82	8879.71	59181.75
std	41.77	42.748	604.44	3097.54	25269.37
min	814	249	8002	3190	15667
25%	881	268	9066	6411	39998.50
50%	901	290	9475	7667	51416
75%	937	335	9965	11073.75	133004
max	992	397	10690	18012.50	162675

Finally, the REG data contains 1847 rows with 8 columns which are labeled as: *YMD*, *Month*, *WeekDay*, *NbHeuresVehYMD*, *NbVehiculesActifYM*, *nbStationsActivesYM*, *nbUsagersActifsYM* and *distTotaleKmReservation*, for analyzing the consumed hours and mileage as responses are ready to serve in this study ("*YMD*" refers to year).

4.1.3 Convert Categorical Variables to Dummy Variables

After cleaning the data and removing the uninformative observations, the next step is to get rid of non-numeric variables for facilitating the data to feed some libraries in Python and models such as multiple regression. Categorical or non-numeric variables can be converted to a numeric dummy variable with two values, zero or one.

In the previous chapter weather observations that are in range of -25 to 30 are converted, into zero and one. In this case, the values in the *WeatherMeannTemp* column are encoded by two levels of possible binary variables. If the values in the *WeatherMeannTemp* column obtain absolute zero or subzero, then cold weather is considered, and takes a value of zero. On the other hand, weather is assumed desirable if the value in *WeatherMeannTemp* is more than zero. The desirability of weather is marked by the dummy variable one. Holiday information dataset is filled by Yes and No. The dummy variables in this dataset are described by 0 instead of No or non-holiday. Moreover, 1 indicates holiday or Yes in the related dataset.

4.1.4 Test Samples

Combination of three datasets under this study includes 10 features with 1847 rows through October 2010 to the end of January 2016. As explained before, this dataset is time dependent. Therefore, data must be divided into different samples with respect to date. Therefore, by applying the forward-changing technique, there are 4 folds as shown in Figure 4.2. Each blue-highlighted year is used to train the models and yellow-marked years are employed to evaluate the performance of created models.

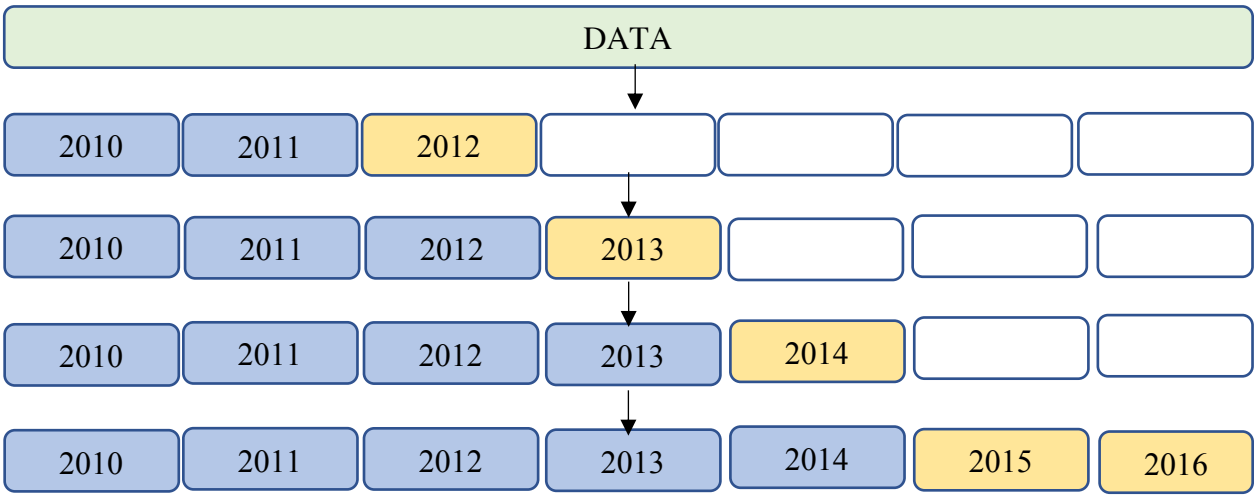

Figure 4-2: Split data to four folds as training and test sets.

4.1.5 Data Visualization

As mentioned at beginning of this chapter, all results in this chapter are relevant to consumed hours as response variable under the study related to REG dataset. Figure 4.3 indicates the density plot of the continues variables. Based on the shapes of density plots of each variable, we conclude that all variables are distributed normally. Furthermore, this figure contains the scatter plots which show the relationship between variables. Discovering patterns between the variables are used in multiple regression model. For example, third row most right-hand plot reveals no linear dependency between the number of stations and users. Moreover, the first row in that figure illustrates the scatter plot of independent variables respected to the consumed hours.

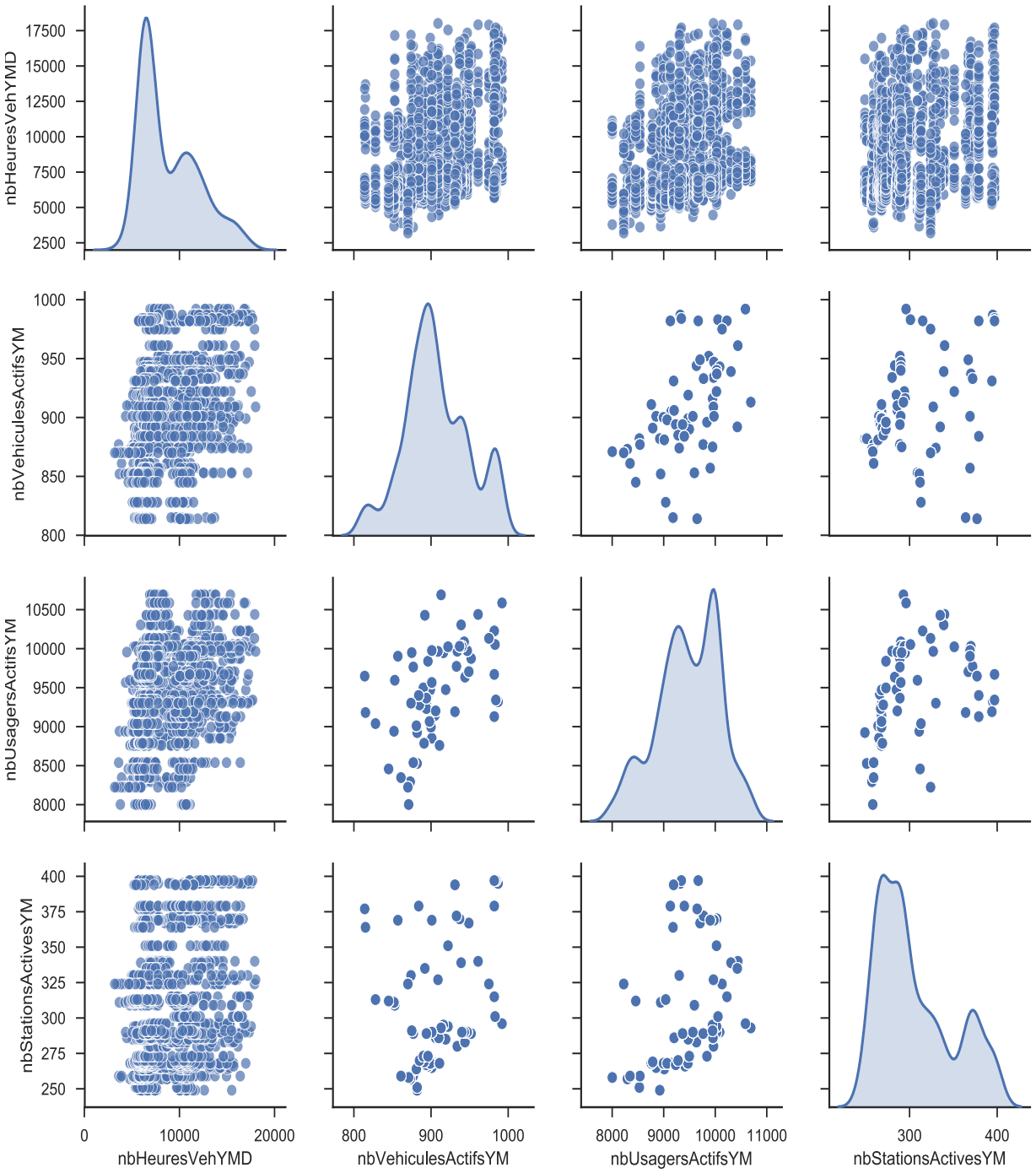

Figure 4-3: Distribution of single variable and scatter plots to show the relationship between variables (REG dataset)

Figure 4.4 shows the correlation between the response value and other factors. Examining the correlation led to removal of the features which are irrelevant to the response value. The results of Pearson method for computing the correlation between variables are shown in the following figure. It is clear that the consumed hours variable is strongly correlated with holiday time and week day with the correlation coefficients of 0.75 and 0.59, respectively.

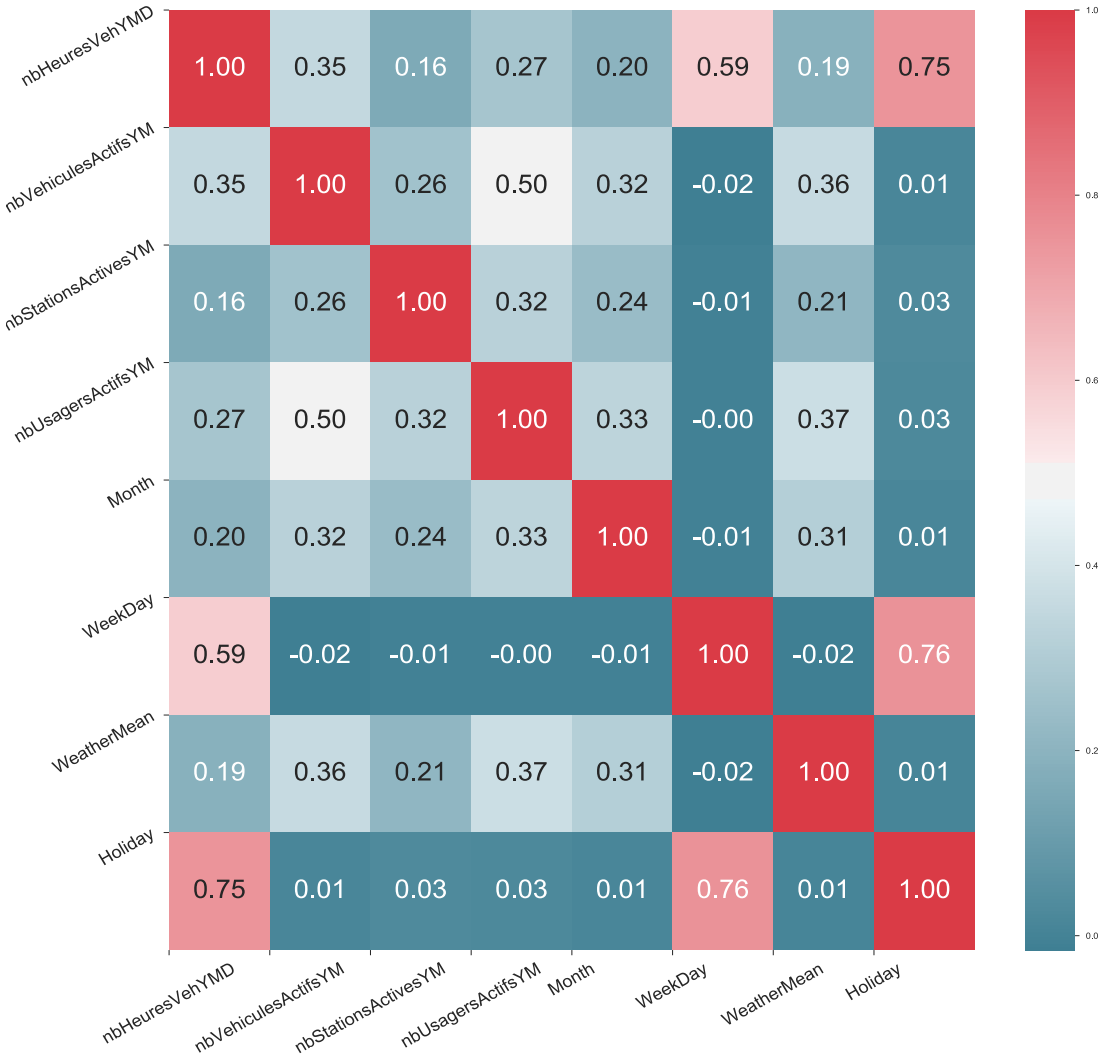

Figure 4-4: Correlation matrix between variables (REG dataset)

A negative reported correlation means that if the value of one variable increases, the other variable decreases. But all of the negative values in this figure are close to zero. It seems that there is no

relationship between those variables. For instance, there is no meaningful relationship between total number of vehicles, stations, and even total users with week day and holiday, because stations and vehicles that the system serves, are typically fixed within a month. Besides, members in this system should register for at least one month to use the regular service. Therefore, there is no significant correlation between week day and users. Moreover, between the number of vehicles, users and stations, there is a significant relationship.

Considering the correlation results, the response has a high correlation with week day. The following box plots are designed to discover the pattern on different days and holiday time in Montreal (Figure 4.5 & 4.6).

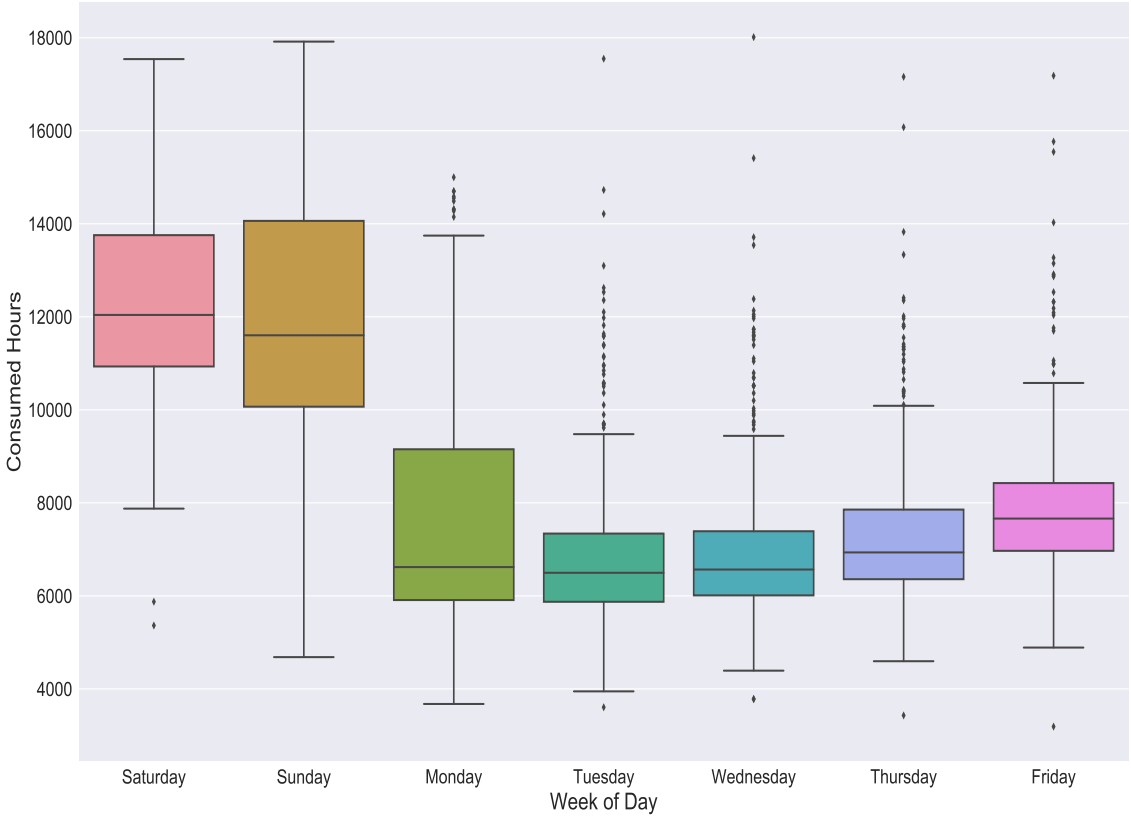

Figure 4-5: Box plots of consumed hours by day of week (REG dataset)

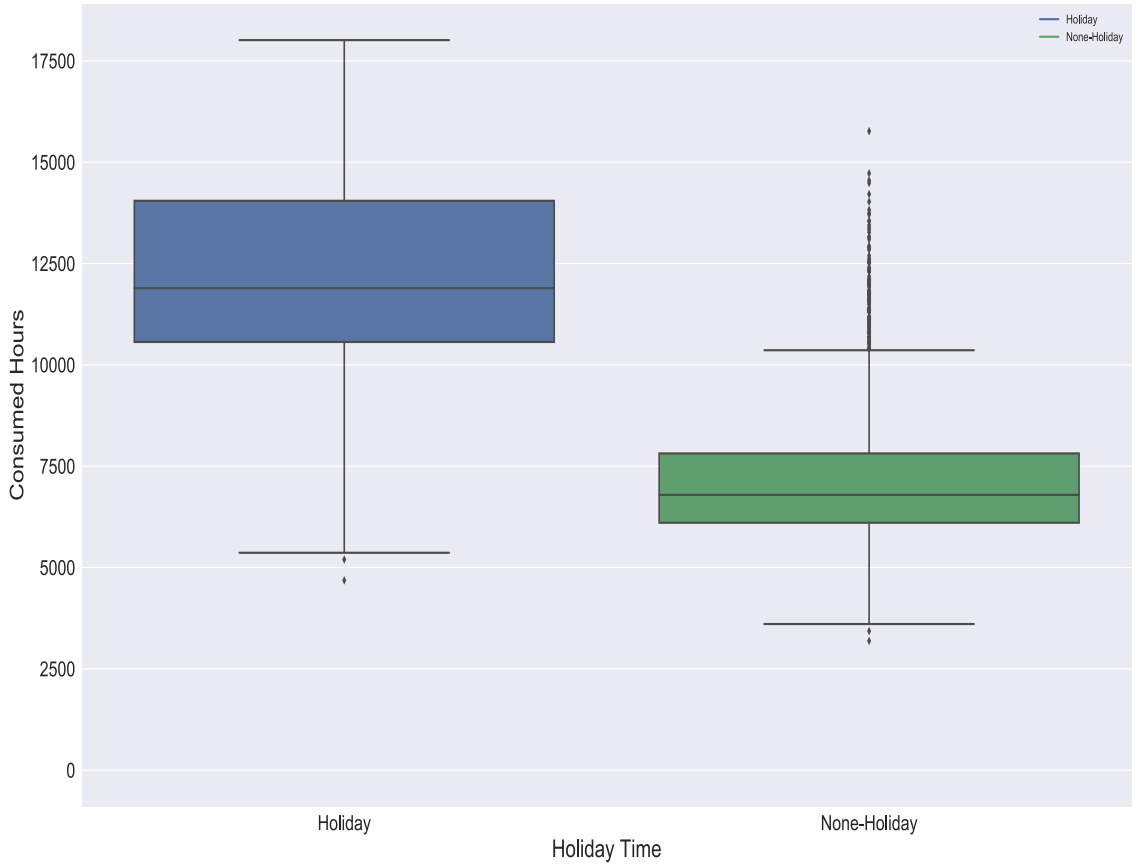

Figure 4-6: Box plots of consumed hours by holiday time (REG dataset)

The Figure shows that the consumed hours ascend significantly on Saturday and Sunday whose medians are more than 11000 hours. In the other days, Monday to Thursday, the median is less than 8000 hours per day.

Figure 4.6 displays the high consumption of station-based service on holiday. Holiday is depicted by Yes in box plot. In this case the peak is in holiday time, where the median is more than 11000 hours, which is similar to the result of the first plot. Interestingly in both plots, holidays have significant influence on the response variable in this project, possibly because during holidays users are more attracted to use this service.

The amount of consumed hours by month is illustrated in the following box plots. Figure 4.7 illustrates that the range of used hours is not the same every month. For instance, in July and August the range of used regular service is drastically more than the other months. The indicated median is more than 10000 hours per month in the mentioned months. The reason might be the summer season, when people are apart from education or able to reserve and use the vehicles in this service for longer time. Moreover, Figure 4.8 shows the number of used hours when the weather average is more than zero or desirable weather versus sub-zero or bad weather. As expected from the Figure 4.7, when the weather is good, the consumption of the vehicles in this service is significantly higher rather the bad weather. It can be hypothesized that members of Communauto use this service more in good weather and holidays.

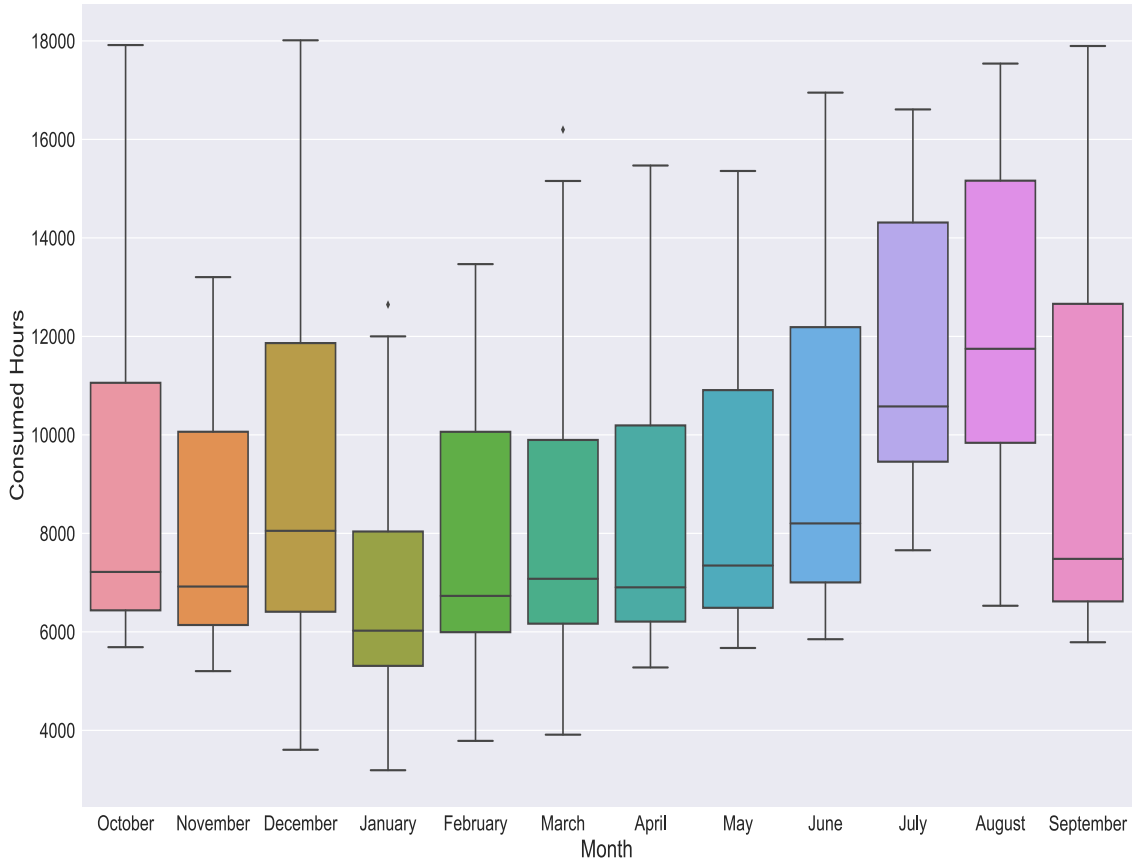

Figure 4-7: Box plots of consumed hours by month (REG dataset)

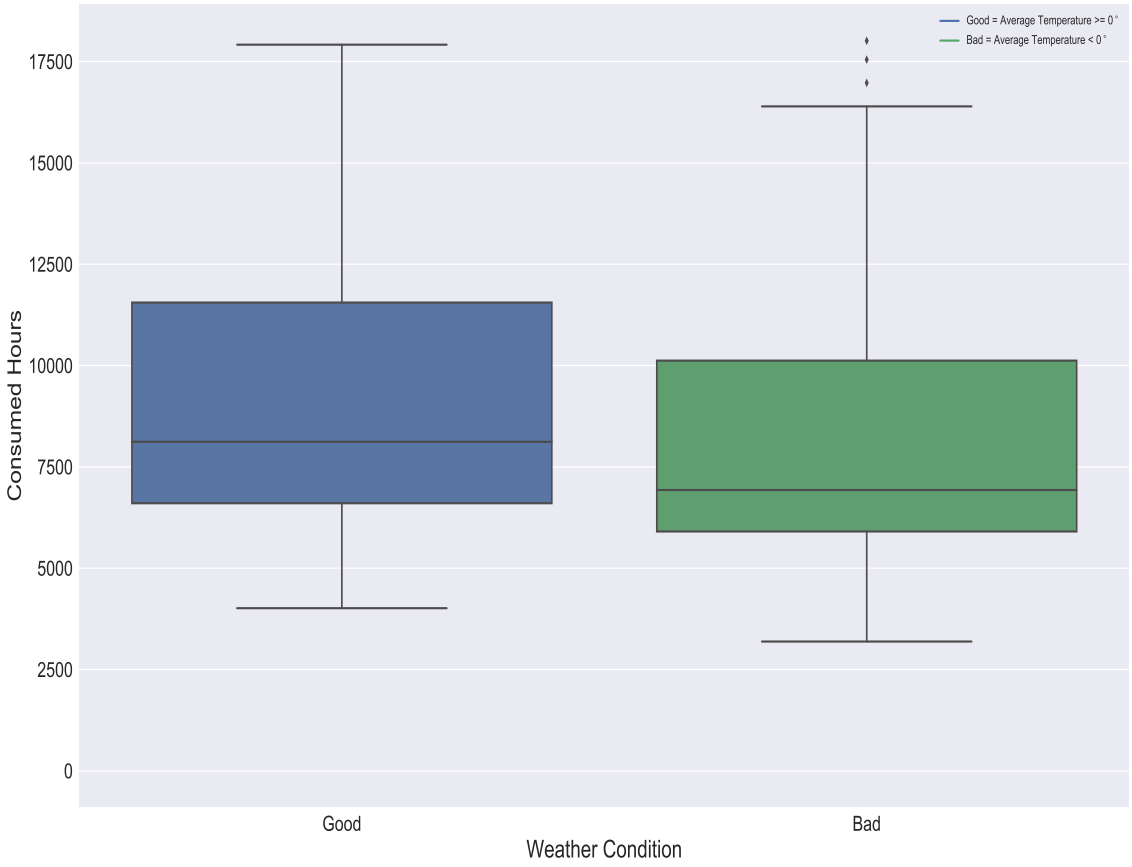

Figure 4-8: Box plots of consumed hours by weather conditions (REG dataset)

4.2 Experiments

In this stage, supervised learning models are built to forecast the consumed hours within REG or regular service data of Communauto operator by using multiple regression, regression tree, random forests, gradient boosting, LSTM and GRUs recurrent neural networks algorithms. These learning methods are non-linear model which are able to capture none-linear feature behavior. Non-linear models can be applied to a vast set of problems.

The training set which is generated in part of data splitting, is used to train the algorithm to build the models, and validated by using the rest, test set. Then the performance of all models is measured with RMSE metrics.

Similarly, in this experiment, all models are applied to AUM data (Auto-Mobile service) for forecasting the consumed hours and mileage. Moreover, the mileage as another target of this thesis in REG data is also investigated as part of the experiment.

It is worth noting that, in the structure of all models, REG data with the following features are considered:

- ***YMD*:** Year
- ***Month*:** The month in each year which is shown by 0 to 12
- ***WeekDay*:** The day of the week which is shown by 0 to 6
- ***nbVehiculesActifsYM*:** The number of available vehicles in each month
- ***nbStationsActivesYM*:** The number of available stations in each month
- ***nbUsagersActifsYM*:** The number of active users in each month

Thereafter, holiday and weather data will be employed to construct the forecasting models. In all models, the year is considered as index in the process, and month and week day are considered as object.

4.2.1 Experiment 1 : Multiple Regression

Based on the REG dataset, six variables as features are involved in forecasting the consumed hours as response of this study. The goal of this step is fitting a model for forecasting the consumed hours in different months.

In the first step, the additive assumption should be removed. It means that the effect of changes in one of the features in data on the desired response, is independent from the other features. Considering scatter plots and correlation in Figures 4.3 and 4.4, it is concluded that all the factors are associated with consumed hours. Significant correlations in REG data are related to the number of available vehicles and users (≈ 0.5), and the number of available stations and users (≈ 0.33). Inclusion of the interaction of those variables is examined separately in the model.

Initially, the model is trained by training set. Moreover, the interaction between the number of available vehicles and users is involved in the model. However, the associated p-value with the coefficient estimate of involved interaction, is 0.654, which is large, more than the considered

p-value = 0.05. It seems that there is no statistical evidence between consumed hours and interaction between those variables.

Afterwards, the model is examined with interaction between the number of available stations and users. The p-value associated to their coefficients is not significant in the model. Therefore, this suggests that it would be appropriate to remove the number of available stations and users and add their interaction term in the model to forecast the consumed hours.

The obtained fitted model takes the following form (4.1):

$$nbHeuresVehYMD \approx \beta_0 + \beta_1 \times nbVehiculesActifsYM + \beta_2 \times (nbStationsActivesYM * nbUsagersActifsYM) + \beta_3 \times Month + \beta_4 \times WeekDay \quad (4.1)$$

The results of coefficient from the fitted model on the training set are shown in Appendix 1.1. The results in the following table strongly suggest that the model which contains interaction between the number of available stations and users, is superior to the model that consists the interaction between the number of available vehicles and users. Moreover, all p-values associated to variables and interaction term are statistically significant. This indicates the evidence for rejecting the null hypothesis $H_0: \beta_p = 0$.

The reported R-squared by the model is around 77%. Thereafter, the trained model is used to forecast the response variable on the test set. In this way, root mean squared error is considered as measure performance of the model.

In the other phase, holiday and weather dataset are involved in fitting model for forecasting the consumed hours as response variable. Firstly, it is hypothesised that the variable in both data does not have significant influence on the model performance. The model is set by adding the average weather and holiday (4.2). The results are brought in Appendix 1.2.

$$nbHeuresVehYMD \approx \beta_0 + \beta_1 \times nbVehiculesActifsYM + \beta_2 \times (nbStationsActivesYM \times nbUsagersActifsY M) + \beta_3 \times Month + \beta_4 \times WeekDay + \beta_5 \times WeatherMean + \beta_6 \times Holiday \quad (4.2)$$

As indicated in the table, R-squared is 83%. The results of R-squared deduced that the model is improved by adding holiday and weather conditions.

Figure 4.9 shows the real values versus predicted values of consumed hours as response variable in both models. Left-hand panel indicates the multiple regression model with RGE dataset; while, right-hand panel is related to the model with additive data (holiday and weather). It is clear that the points are close to the solid red line in the second model. It means that forecasting model with all dataset outperforms a model without holiday and weather data which is indicated in RMSE results.

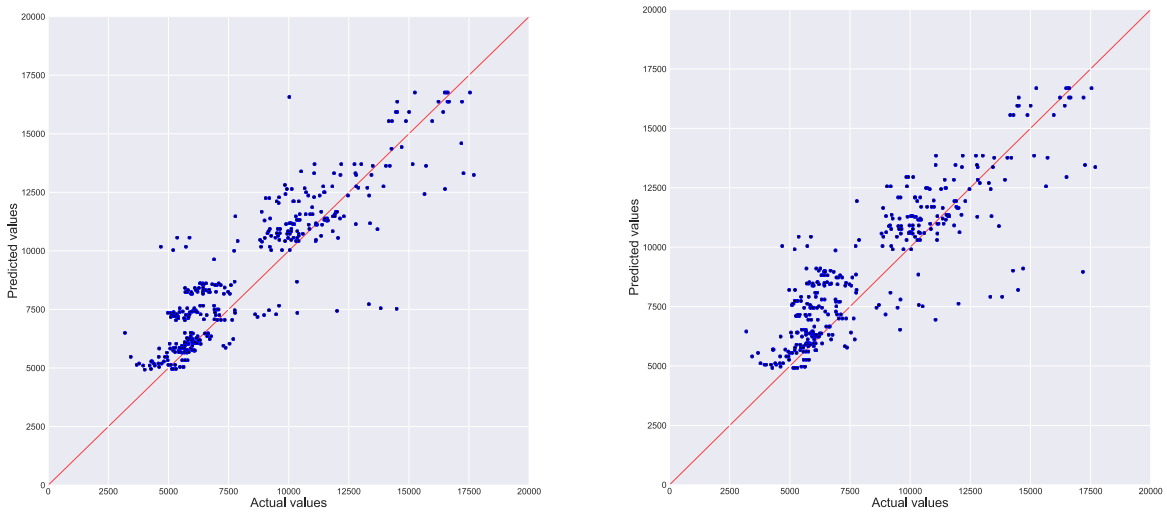

Figure 4-9: Left: The plot of real values vs. predicted values in multiple regression model by REG data. Right: The plot of real values vs. predicted values in the same model by REG data, holiday and weather dataset.

4.2.2 Experiment 2: Regression Tree

Regression tree is fitted to the REG data set. First, the model is trained by training set, and then it is examined by test set. The maximum depth is considered 5 and MSE is applied as indicator to split the data in different branches. The feature labeled as *WeekDay* is chosen for root node by the model. Regression tree algorithms are also trained by the REG data combined with holiday and weather features. The Figure 4.10 depicts the real values versus predicted values of both trees.

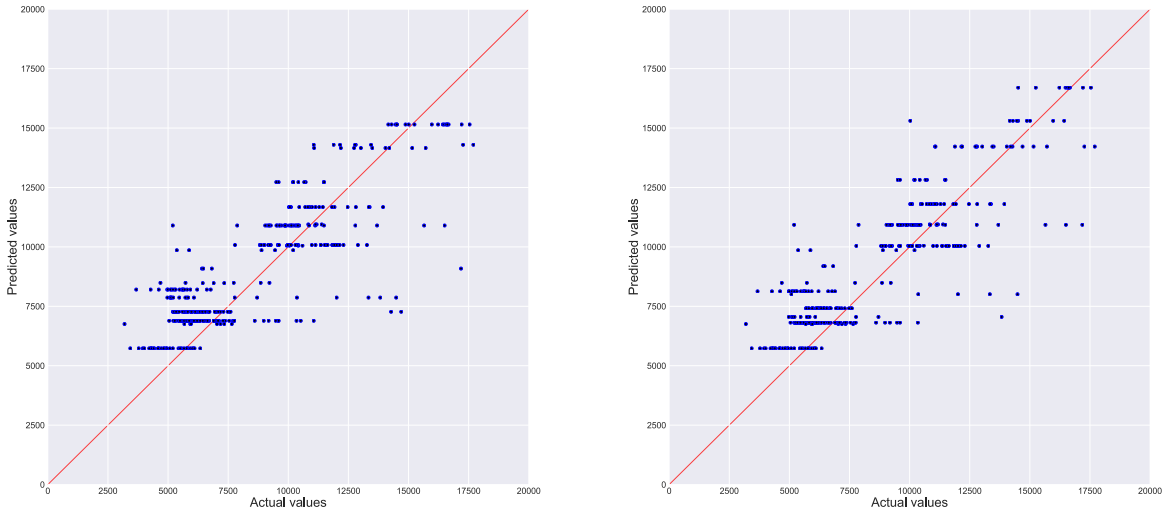

Figure 4-10: Left: The plot of real values vs. predicted values in multiple regression model by REG data. Right: The plot of real values vs. predicted values in the same model by REG data, holiday and weather dataset.

The additional features possibly do not really assist to improve the forecasting model in the plot; while, the results of RMSE indicate that the additional features improve the performance of the model.

4.2.3 Experiment 3: Random Forests

The principal element of tuning random forests algorithm is the number of estimators. In this step, 1000 estimators are founded to produce the best forecasting results. Moreover, MSE is used as an indicator to split for training the random forests algorithm. Model is trained by REG data, and in the other phase, holiday and weather are joined to the data. The results of predicted values versus actual values of consumed hours are shown in Figure 4.11. The left-hand panel indicates the random forests model with only REG data with six features, and the right-hand panel shows the same model with additional data, holiday and weather. By comparing the plots, the additional features do not may not assist to improve the forecasting model in forecasting action; while, there is a significant difference in their root mean squared errors.

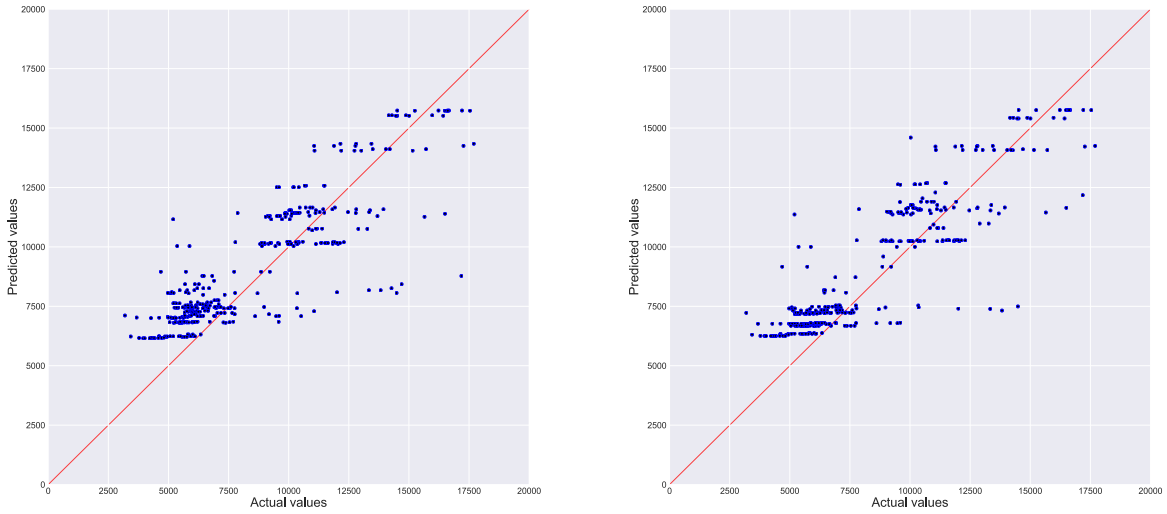

Figure 4-11: Left: The plot of real values vs. predicted values in multiple regression model by REG data. Right: The plot of real values vs. predicted values in the same model by REG data, holiday and weather dataset.

Moreover, Figure 4.12 depicts the important independent variables or features in training process of the trees in this technique. Left-hand panel shows the importance variables of REG data in conducting the model. As seen, the highest bar chart in left-hand panel, labeled as *WeekDay*, has the most impact on the forecasting process. In contrast, the number of active stations has the lowest impact in the model.

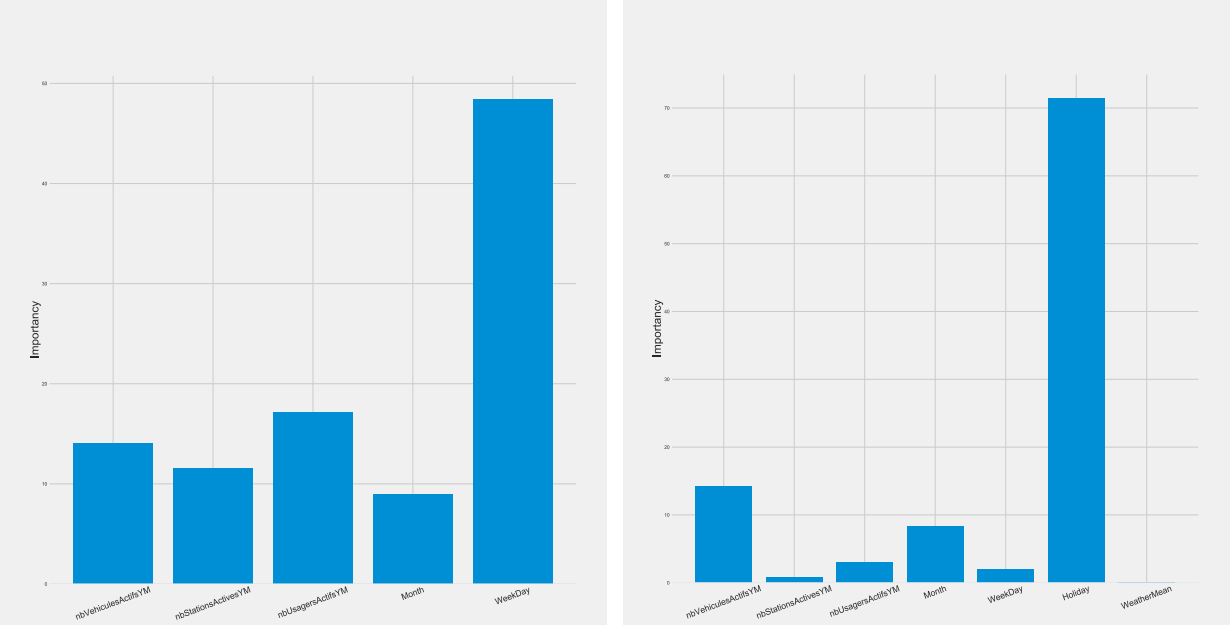

Figure 4-12: Left: Bar chart of importance variable in random forests model with REG data. Right: Bar chart of importance variable in random forests model with REG data and additive features.

In the right-hand panel, two more variables are added to the structure of the model. The feature labeled as *WeatherMean* has no influence, while holiday feature is the most important variable. The *WeatherMean* variable is excluded in the model, but RMSE is not changed. So that I decided to keep this feature in the model.

4.2.4 Experiment 4: Gradient Boosting

For training gradient boosting, the baseline is established with MSE function as the indicator to split, and least squares as loss function. Maximum depth (δ) is considered 5, because it produces the best quality forecasts. Learning rate or shrinkage (α) is scheduled 0.1. Gradient boosting is trained by two groups of data, with and without holiday and weather variables.

In Figure 4.13, the left-hand panel is related to predicted values with only REG dataset and in the right, holiday and weather features are added. The right-hand plot looks like to be more coherent around the solid red line.

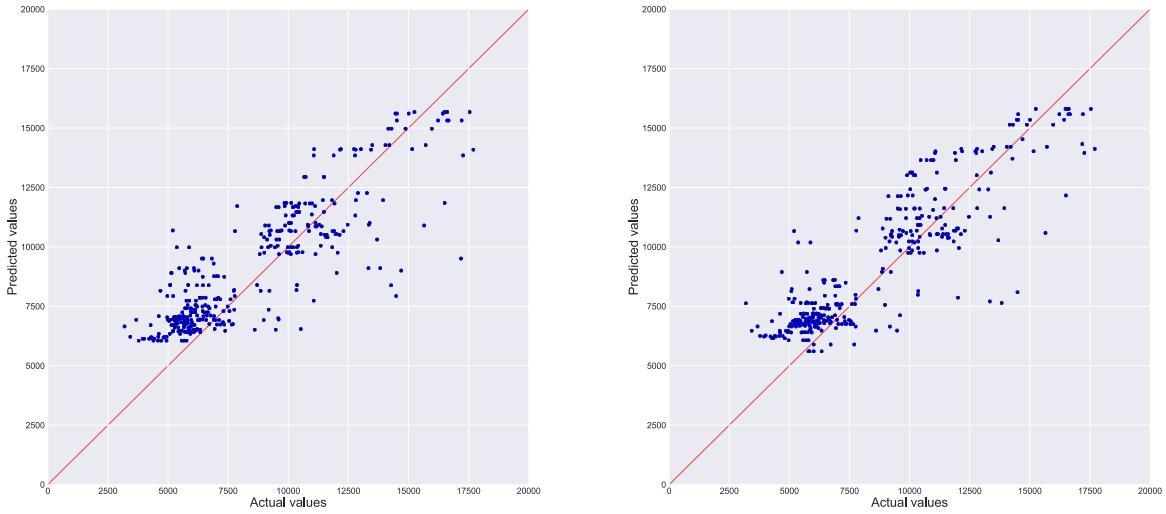

Figure 4-13: Left: The plot of real values vs. predicted values in multiple regression model by REG data. Right: The plot of real values vs. predicted values in the same model by REG data, holiday and weather dataset.

4.2.5 Experiment 5: Long-Short Term Memory RNNs

As mentioned in section 3.3.4, the data were normalized for this algorithm, from zero to one. The loss function in this algorithm was employed by MSE with rmsprop for an optimiser, and tanh output layer.

Hidden layer is considered 5 and epoch is scheduled 1000. The history plot (Figure 4.14) of loss in training and test set shows that when the epoch is less than 200, the error received from loss function in the algorithm decreases dramatically; while, with 1000 epoch, the line goes stable. Therefore, 1000 is set for epoch in this algorithm.

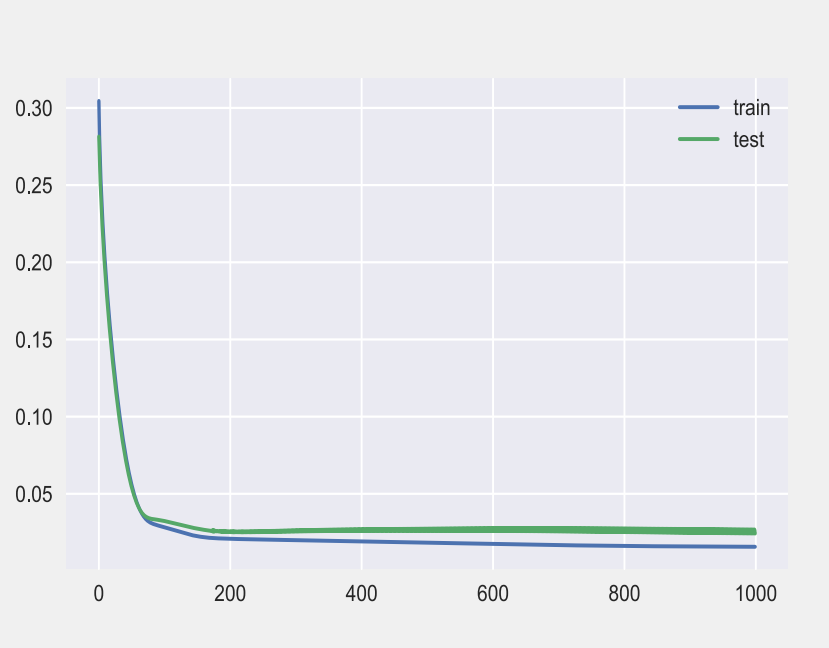

Figure 4-14: History plot of loss function

After the process of creating the forecasting model, the normalized data should be converted to their initial origin for computing RMSE as the evaluation metric in this study.

Figure 4.15 shows the actual and predicted values in both statements, by REG data and additive data (holiday and weather).

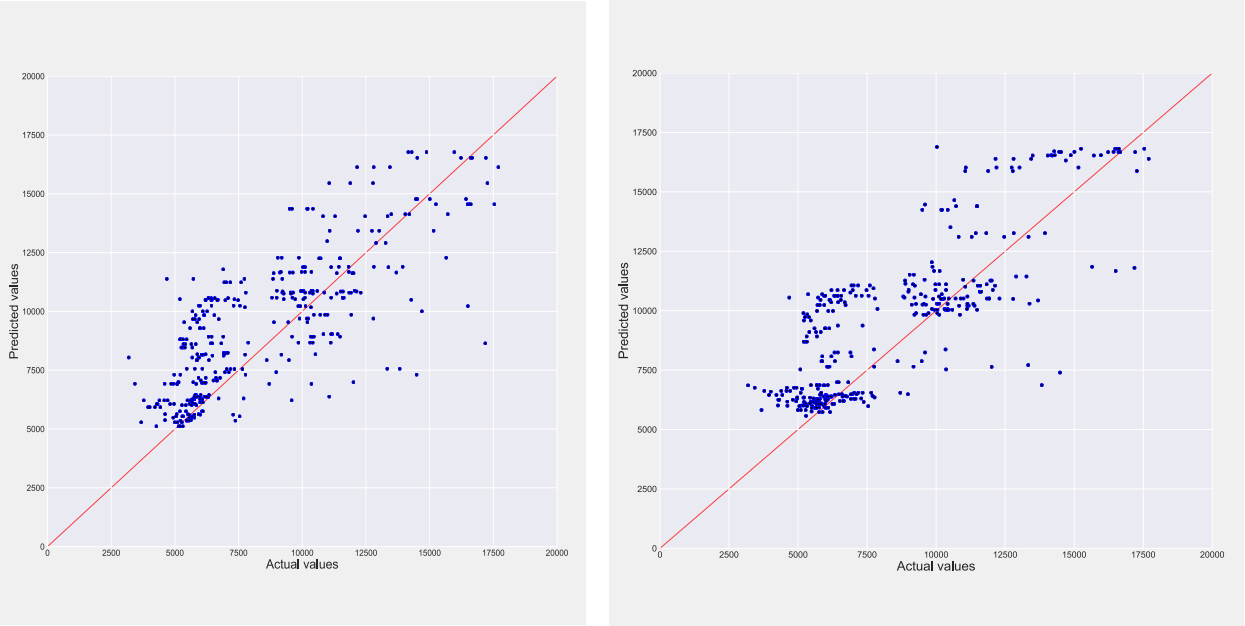

Figure 4-15: Left: The plot of real values vs. predicted values in multiple regression model by REG data. Right: The plot of real values vs. predicted values in the same model by REG data, holiday and weather dataset.

4.2.6 Gated Recurrent Units RNNs

Similar to LSTM model, the observations were normalized between zero and one. The hidden layer is considered 10 with 1000 epoch. The history plot in Figure 4.16 depicts the stability of results of loss function (MSE) when epoch is 1000. In addition, after forecasting process, the normalized data is converted in order to compute RMSE.

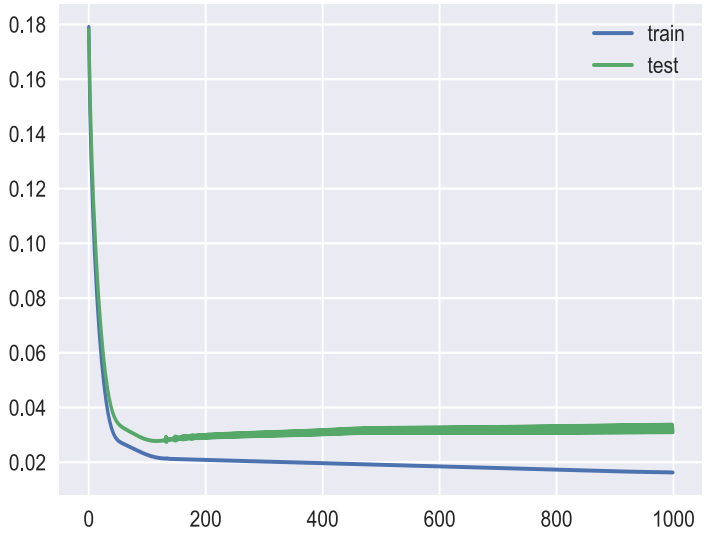

Figure 4-16: History plot of loss function

Moreover, Figure 4.17 produces the scatter plots of real values versus predicted values. It is clearly observed that in the right figure the points tend to be close to red solid line.

Figure 4-17: Left: The plot of real values vs. predicted values in multiple regression model by REG data. Right: The plot of real values vs. predicted values in the same model by REG data, holiday and weather dataset.

4.3 Evaluating Forecasting Models

The obtained results through all the models related to REG data and further REG data with holiday and weather features are designed in the following tables. Therefore, the results of RMSE in the tables represent the performance of the forecasting models. These results are achieved by the trained models such as multiple regression, regression tree, random forests, gradient boosting, LSTM and GRUs recurrent neural networks, on 4-folds produced test sets by forward-chaining technique. The models that give the lowest level of test RMSE are highlighted. Most of the models report MSE as the distance between real and predicted values; in order to obtain the RMSE, a root must be taken from the reported quantities.

Figure 4.18 presents the performance of forecasting models, which have been conducted on the REG data respected to consumed hours. It can be clearly interpreted from Table 4.2 and box-whisker plot of RMSE that gradient boosting algorithm with the lowest error rate in the forecasting of response variable outperforms the other models. Considering the results of the table, the poor estimates are produced by LSTM and GRUs; whereas, LSTM outperforms GRUs, which was

expected. Mostly, GRUs train faster and perform better than LSTM but for smaller dataset, as mentioned in section 3.5.5.2.

Table 4-2: The experiment results obtained from forecasting models by using REG data

Model \ RMSE	Fold 1 (01.01.2012 to 31.12.2012)	Fold 2 (01.01.2013 to 31.12.2013)	Fold 3 (01.01.2014 to 31.12.2014)	Fold 4 (01.01.2015 to 31.01.2016)	Average
Multiple Regression	1380.82	1470.53	1690.03	1779.22	1580.15
Regression Tree	1503.83	1473.10	1503.83	1870.17	1587.73
Random Forests	1438.56	1386.28	1440.48	1834.58	1524.98
Gradient Boosting	**1300.86**	**1319.82**	**1354.35**	**1774.88**	**1437.48**
LSTM RNNs	2058.22	2257.05	1890.25	2430.69	2159.05
GRU RNNs	2131.681	2148.182	1973.09	2607.62	2215.14

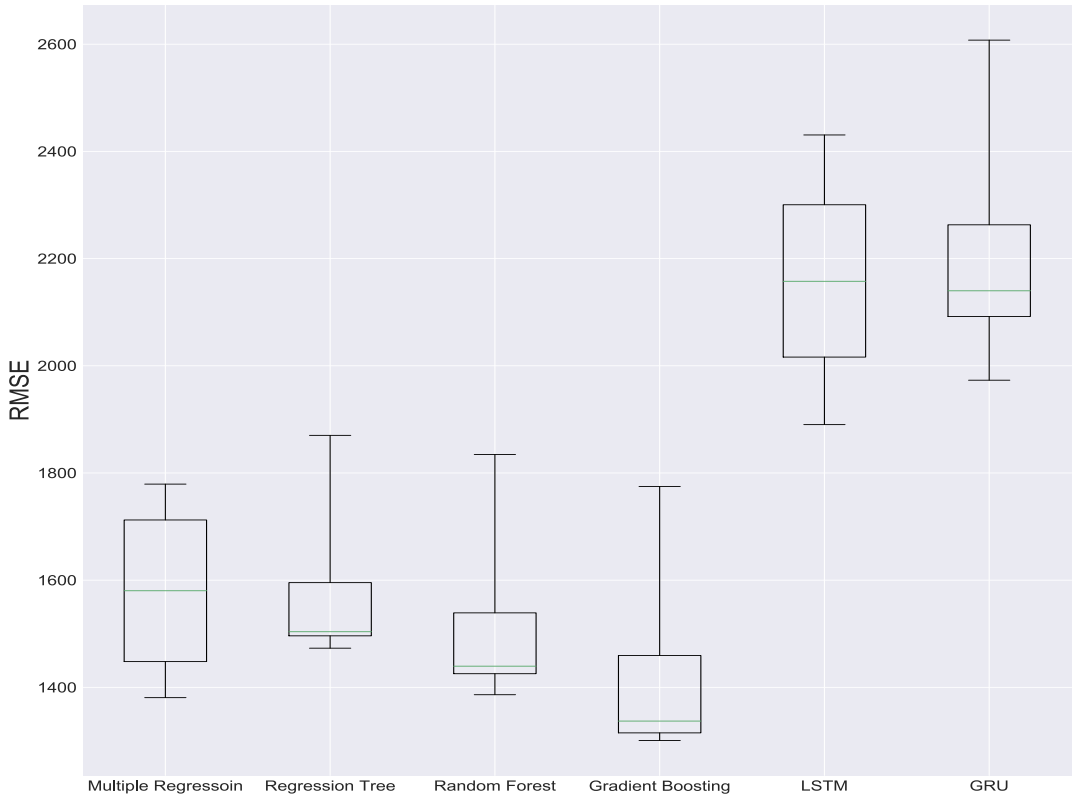

Figure 4-18: Comparison of the error measurement RMSE of models on REG data by box-whisker plot.

Consequently, other experiments have been performed on REG data combined with holiday and weather data. Table 4.3 and Figure 4.19 show the performance of forecasting models after engaging the additional features. Likewise, in this step, gradient boosting has the lowest RMSE and both neural networks algorithms have the worst forecasting of response values.

Table 4-3: The experiment results obtained from forecasting models by using REG data combined with holiday and weather data

Model \ RMSE	Fold 1 (01.01.2012 to 31.12.2012)	Fold 2 (01.01.2013 to 31.12.2013)	Fold 3 (01.01.2014 to 31.12.2014)	Fold 4 (01.01.2015 to 31.01.2016)	Average
Multiple Regression	1235.42	1414.89	1477.69	1631.12	1439.78
Regression Tree	1356.76	1268.31	1275.45	1723.37	1405.97
Random Forests	136.23	1151.82	1189.05	1689	1291.53
Gradient Boosting	**1083.81**	**1034.46**	**1049.14**	**1680.42**	**1211.96**
LSTM RNNs	1678.57	1851.60	1777.27	2314.04	1905.04
GRU RNNs	1803.89	1775.99	1735.67	2436.522	1938.02

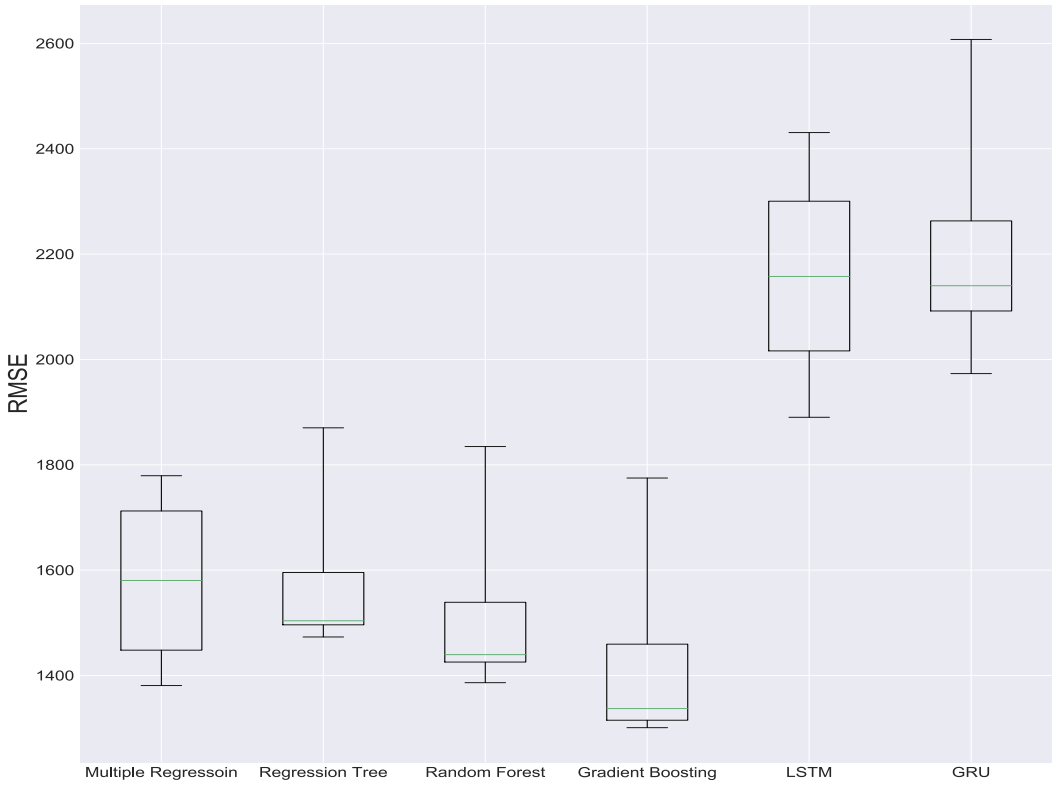

Figure 4-19: Comparison of the error measurement RMSE of models on REG data combined with holiday and weather data by box-whisker plot

Table 4.4 shows the results of RMSE when holiday and weather data were combined to the REG data. The purpose of this table is to discover differences in how the models handle two groups of data compared together. The most noticeable change in these experiments is the significant decrease in the RMSE of each model after adding holiday and weather information. Generally, all the trained models by second group of data achieved better forecasting performance than the first group. These additional features brought more information in the forecasting process.

Table 4-4: Experiment results overview by applying forecasting models

Data / Model	REG	REG + Holiday + Weather
Multiple Regression	1580.15	1439.78
Regression Tree	1587.73	1405.97
Random Forests	1524.98	1291.53
Gradient Boosting	**1437.48**	**1211.96**
LSTM RNNs	2159.05	1905.04
GRUs RNNs	2215.14	1938.02

CHAPTER 5 CONCLUSION AND RECOMMANDATIONS

The main goal of this thesis has been to develop multiple supervised machine learning algorithms, based on historical dataset. The data was collected by Communauto carsharing operator providing regular and free-floating services. Therefore, this project processed the regular service data and free-floating service data individually. The desired responses under this study are consumed hours and mileage. I monitored these responses in different months in respect to type of day in a week. For instance, I concluded that regular or station-based service is used most in July and August. Moreover, members use this service more on Saturdays and Sundays, which indicates that the usage rate of this service on weekends is more aggregated than workdays.

The usage rate of free-floating service or Auto-Mobile vehicles in different months and days of a week is almost the same, with a non-significant difference. The box plots of the results are attached in Appendix 2.

All transactions are implemented by application and online service. Therefore, some observations cannot be trusted, because the system may be updated at a time, or it may be interrupted. Therefore, the observations with abnormal behavior and incomplete data were discarded from the initial dataset. Forecasting models such as multiple regression, regression tree, random forests, gradient boosting, LSTM and GRUs recurrent neural networks were applied on Communauto dataset.

In the next phase, I engaged other features as additive data, such as Quebec province holidays and weather conditions in Montreal. Considering the designed box plot related to regular service dataset, I found that when the weather is good, or at a holiday time, users of regular service are more eager to use this type of service. While, the box-plot in Appendix 3 shows that the rate of used hours and mileage of free-floating service is not under the influence of holiday and weather conditions.

The most important goal of this thesis has been to evaluate the performance of applied supervised learning models to forecast the desired responses. Moreover, another purpose of this work has been to investigate the effect of additional factors on performance of models. The forecasting process of consumed hours as response related to REG data or regular service is discussed in chapters 3 and 4.

The results of hours and mileage forecasting related to AUM or free-floating and regular services are explained below separately.

5.1 Regular Service

As it is shown in chapter 4 and the attached results of REG variables respect to mileage in Appendix 2, trained gradient boosting has the lowest RMSE compared to other models. Therefore, gradient boosting outperforms the other algorithms. Gradient boosting algorithm does not need to find the relationship between variables, and this model is more robust. The algorithm is inherently set to find the relationship. Moreover, the worst results are related to LSTM and GRUs recurrent neural networks. As expected, LSTM had better performance than GRUs, because GRUs is appropriate for problems that require lower memory or smaller dataset. Concisely, statistical methods such as multiple regression and tree-based models outperformed artificial neural network models.

The models were applied on two groups of data: without holiday and weather variables; with holiday and weather variables. In both, it was clearly observed that the additional features helped the enhancement of the performance of forecasting models. Therefore, the consumed hours and mileage in regular service dependent on holiday times and weather conditions.

5.2 Free-Floating Service

In free-floating service, the data is split to October 2013 to January 2014 as training set, and January 2015 to January 2016 as test set. The results of forecasting models are shown in Appendix 3. The additive features improved the models, but not with a significant difference.

Multiple regression outperformed the other models. Gradient boosting in this structure of data did not perform well. It may be mainly because the relationship between the variables is linear, while tree-based models are suitable when the relationship between variables is non-linear.

5.3 Future Directions

There are a lot of independent variables which can help to have an accurate forecasting model. Weather variable in this project is classified into good and bad weather. Use of complete information about weather and temperature, such as precipitation rates (rain or snow), and sunny or cloudy weather conditions, can have an effective impact on performance of models.

The data in this thesis were collected considering day, month and year. Monitoring the usage trend by hours can also be helpful in the future forecasting process.

BIBLIOGRAPHY

Abhijit Mondal. (2017). What is the role of activation function in neural networks? Retrieved from http://www.stokastik.in/machine-learning-interview-questions-and-answers-part-i/

Alpaydin, E. (2009). *Introduction to machine learning*: MIT press.

Augustin, S., Muntaner, L., Altamirano, J. T., González, A., Saperas, E., Dot, J., . . . Esteban, R. (2009). Predicting early mortality after acute variceal hemorrhage based on classification and regression tree analysis. *Clinical Gastroenterology and Hepatology, 7*(12), 1347-1354.

Bean, J. P. (1982). Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in Higher Education, 17*(4), 291-320.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks, 5*(2), 157-166.

Bergmeir, C., & Benitez, J. M. (2011). *Forecaster performance evaluation with cross-validation and variants.* Paper presented at the 2011 11th International Conference on Intelligent Systems Design and Applications.

Bingham, N. H. (2006). Heroic periods. *Mathématiques et sciences humaines. Mathematics and social sciences*(176), 31-42.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth Int. *Group, 37*(15), 237-251.

Chen, P.-C., Hsieh, H.-Y., Sigalingging, X. K., Chen, Y.-R., & Leu, J.-S. (2017). *Prediction of Station Level Demand in a Bike Sharing System Using Recurrent Neural Networks.* Paper presented at the Vehicular Technology Conference (VTC Spring), 2017 IEEE 85th.

Chen, Y., Luo, T., Liu, S., Zhang, S., He, L., Wang, J., . . . Sun, N. (2014). *Dadiannao: A machine-learning supercomputer.* Paper presented at the Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Communauto Inc. (2019). Find a vehicle. Retrieved from https://www.communauto.com/index_en.html

Danny Mathew. (2018). Intro to recurrent neural networks lstm | gru. Retrieved from https://www.kaggle.com/honeysingh/intro-to-recurrent-neural-networks-lstm-gru

Das, K., & Behera, R. N. (2017). A survey on machine learning: concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering, 5*(2), 1301-1309.

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology, 81*(11), 3178-3192.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning *The elements of statistical learning* (pp. 485-585): Springer.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Ji, S., Cherry, C. R., Han, L. D., & Jordan, D. A. (2014). Electric bike sharing: simulation of user demand and system availability. *Journal of Cleaner Production, 85*, 250-257.

Joelsson, S. R., Benediktsson, J. A., & Sveinsson, J. R. (2005). *Random forest classifiers for hyperspectral data.* Paper presented at the Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. Proceedings. 2005 IEEE International.

Kang, D., Lv, Y., & Chen, Y.-y. (2017). *Short-term traffic flow prediction with LSTM recurrent neural network.* Paper presented at the Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*: John Wiley & Sons.

Karalic, A., & Cestnik, B. (1991). *The bayesian approach to tree-structured regression.* Paper presented at the Proceedings of ITI.

Katzev, R. (2003). Car sharing: A new approach to urban transportation problems. *Analyses of Social Issues and Public Policy, 3*(1), 65-86.

Klein, C., Foerster, F., Hartnegg, K., & Fischer, B. (2005). Lifespan development of pro-and anti-saccades: multiple regression models for point estimates. *Developmental Brain Research, 160*(2), 113-123.

Kohavi, R., & Quinlan, J. R. (2002). *Data mining tasks and methods: Classification: decision-tree discovery.* Paper presented at the Handbook of data mining and knowledge discovery.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science, 1*(2), 111-117.

Krishnamurti, T., Kishtawal, C., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., . . . Surendran, S. (1999). Improved weather and seasonal climate forecasts from multimodel superensemble. *Science, 285*(5433), 1548-1550.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26): Springer.

Lang, H. (2013). Topics on Applied Mathematical Statistics. *KTH Teknikvetenskap, version 0.97*.

Lapedes, A., & Farber, R. (1987). *Nonlinear signal processing using neural networks: Prediction and system modelling*. Retrieved from

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Martin, E. W., & Shaheen, S. A. (2011). Greenhouse gas emission impacts of carsharing in North America. *IEEE Transactions on intelligent transportation systems, 12*(4), 1074-1086.

Mohri, M., Talwalkar, A., & Rostamizadeh, A. (2012). *Foundations of machine learning (adaptive computation and machine learning series)*: Mit Press Cambridge, MA.

Murray, A. T., Davis, R., Stimson, R. J., & Ferreira, L. (1998). Public transportation access. *Transportation Research Part D: Transport and Environment, 3*(5), 319-328.

Osborne, J. W. (2000). Prediction in multiple regression. *Practical Assessment, Research & Evaluation, 7*(2), 1-9.

Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. (2007). Random forest models to predict aqueous solubility. *Journal of chemical information and modeling, 47*(1), 150-158.

Persson, C., Bacher, P., Shiga, T., & Madsen, H. (2017). Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy, 150*, 423-436.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*: Malaysia; Pearson Education Limited.

Scheffer, J. (2002). Dealing with missing data.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

Shaheen, S. A., & Cohen, A. P. (2007). Growth in worldwide carsharing: An international comparison. *Transportation Research Record, 1992*(1), 81-89.

Sioui, L., Morency, C., & Trépanier, M. (2013). How carsharing affects the travel behavior of households: a case study of Montréal, Canada. *International Journal of Sustainable Transportation, 7*(1), 52-69.

Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic zoology, 35*(4), 627-632.

Steininger, K., Vogl, C., & Zettl, R. (1996). Car-sharing organizations: The size of the market segment and revealed change in mobility behavior. *Transport Policy, 3*(4), 177-185.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods, 14*(4), 323.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Zani, S., Riani, M., & Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis, 28*(3), 257-270.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting, 14*(1), 35-62.

APPENDIX A- THE RESULTS OF MULTIPLE REGRESSION ON "REG" DATA

Table A-1:The table of results of multiple regression on REG data

Dep. Variable:	nbHeuresVehYMD	**R-squared:**	0.763
Model:	OLS	**Adj. R-squared:**	0.760
Method:	Least Squares	**F-statistic:**	246.9
No. Observations:	1478	**Prob (F-statistic):**	0.00
Df Residuals:	1458	**Log-Likelihood:**	-12883.
Df Model:	19	**AIC:**	2.581e+04
		BIC:	2.591e+04

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7739.2823	1472.932	-5.254	0.000	-1.06e+04	-4849.991
Month[T.2]	892.1036	196.979	4.529	0.000	505.712	1278.496
Month[T.3]	1126.8588	194.149	5.804	0.000	746.018	1507.699
Month[T.4]	1002.0813	197.815	5.066	0.000	614.050	1390.113
Month[T.5]	1193.4833	199.125	5.994	0.000	802.882	1584.084
Month[T.6]	1486.4751	206.212	7.208	0.000	1081.972	1890.978
Month[T.7]	3418.2038	206.985	16.514	0.000	3012.183	3824.224
Month[T.8]	3908.4711	206.037	18.970	0.000	3504.310	4312.632
Month[T.9]	999.9145	204.964	4.878	0.000	597.859	1401.970
Month[T.10]	1008.4803	189.532	5.321	0.000	636.697	1380.264
Month[T.11]	480.5026	187.498	2.563	0.010	112.707	848.298

Table A-1: The Table of results of multiple regression on REG data (cont'd and end)

Month[T.12]	2130.3680	186.107	11.447	0.000	1765.302	2495.434
WeekDay[T.1]	-632.0309	141.427	-4.469	0.000	-909.454	-354.608
WeekDay[T.2]	-577.1953	141.269	-4.086	0.000	-854.308	-300.083
WeekDay[T.3]	-288.9316	142.294	-2.031	0.042	-568.054	-9.809
WeekDay[T.4]	761.3729	153.269	4.968	0.000	460.721	1062.025
WeekDay[T.5]	4751.3340	142.444	33.356	0.000	4471.917	5030.751
WeekDay[T.6]	4356.8304	141.262	30.842	0.000	4079.733	4633.928
nbVehiculesActifsYM	12.9519	1.725	7.509	0.000	9.568	16.335
I(nbStationsActivesYM * nbUsagersActifsYM)	0.0008	0.000	7.873	0.000	0.001	0.001

Table A-2: The Table of results of multiple regression on REG data combined with holiday and weather

Dep. Variable:	nbHeuresVehYMD	**R-squared:**	0.834
Model:	OLS	**Adj. R-squared:**	0.832
Method:	Least Squares	**F-statistic:**	348.2
No. Observations:	1478	**Prob (F-statistic):**	0.00
Df Residuals:	1456	**Log-Likelihood:**	-12619.
Df Model:	21	**AIC:**	2.528e+04
		BIC:	2.540e+04

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7945.2662	1233.384	-6.442	0.000	-1.04e+04	-5525.867
Month[T.2]	894.2183	165.054	5.418	0.000	570.449	1217.988
Month[T.3]	1177.6707	166.389	7.078	0.000	851.283	1504.059
Month[T.4]	951.4305	192.131	4.952	0.000	574.547	1328.314
Month[T.5]	1166.0799	196.079	5.947	0.000	781.452	1550.708
Month[T.6]	1505.0959	200.829	7.494	0.000	1111.151	1899.041
Month[T.7]	3494.1436	201.120	17.373	0.000	3099.628	3888.659
Month[T.8]	4094.8734	200.770	20.396	0.000	3701.044	4488.703
Month[T.9]	983.2388	200.144	4.913	0.000	590.637	1375.840
Month[T.10]	984.0238	188.786	5.212	0.000	613.702	1354.345
Month[T.11]	603.2997	170.916	3.530	0.000	268.033	938.567

Table A-2: The Table of results of multiple regression on REG data combined with holiday and weather (cont'd and end)

Month[T.12]	1892.9056	157.031	12.054	0.000	1584.874	2200.937
WeekDay[T.1]	-174.6126	119.840	-1.457	0.145	-409.690	60.464
WeekDay[T.2]	-89.5410	119.891	-0.747	0.455	-324.718	145.636
WeekDay[T.3]	197.1062	120.738	1.633	0.103	-39.732	433.945
WeekDay[T.4]	1220.3182	129.648	9.413	0.000	966.002	1474.635
WeekDay[T.5]	-729.6442	249.993	-2.919	0.004	-1220.030	-239.258
WeekDay[T.6]	-1118.3171	249.376	-4.484	0.000	-1607.491	-629.143
WeatherMean[T.1]	-172.6037	112.985	-1.528	0.001	-394.234	49.027
Holiday[T.1]	6016.4511	241.246	24.939	0.000	5543.225	6489.677
nbVehiculesActifsYM	12.9099	1.444	8.938	0.000	10.077	15.743
I(nbStationsActivesYM * nbUsagersActifsYM)	0.0007	8.65e-05	8.653	0.000	0.001	0.001

APPENDIX B- EXPERIMENT RESULTS OVERVIEW ON "REG" DATASET RESPECTED TO MILEAGE

Table B-1: Experiment results overview by applying forecasting models

Data / Model	REG	REG + Holiday + Weather
Multiple Regression	11507.18	10985.42
Regression Tree	11507.18	12121.33
Random Forests	10511.86	10295.54
Gradient Boosting	**9711.77**	**9244.98**
LSTM RNNs	20343.66	19680.19
GRUs RNNs	21029.15	19650.45

APPENDIX C- "AUM" DATASET (FREE-FLOATING SERVICE) RESPECTED TO CONSUMED HOURS AND MILEAGE

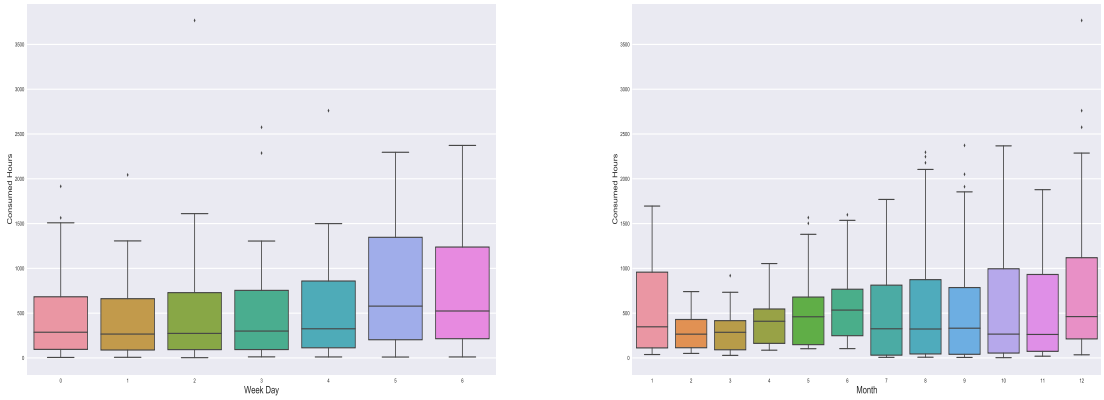

Figure C-1: Box plot of consumed hours respect to week day and month in AUM dataset

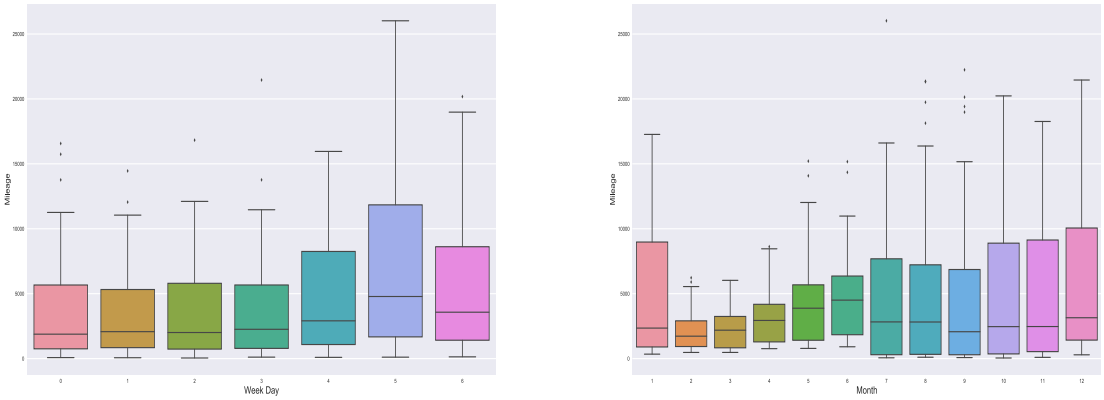

Figure C-2: Box-plot of mileage respect to week day and month in AUM dataset

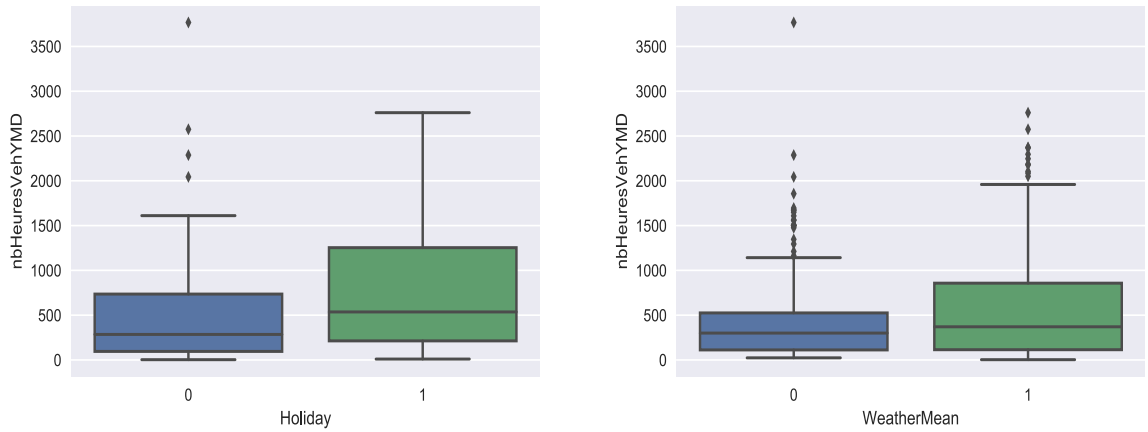

Figure C-3: Box-plot of consumed hours respect to holiday and weather in AUM dataset

Table C-1: Experiment results overview on AUM dataset respected to Consumed hours

Data / Model	REG	AUM + Holiday + Weather
Multiple Regression	**369.65**	**357.78**
Regression Tree	720.04	715.35
Random Forests	669.45	664.71
Gradient Boosting	660.50	649.35
LSTM RNNs	890.20	873.31
GRUs RNNs	899.54	882.83

Table C-2: Experiment results overview on AUM dataset respected to mileage

Data / Model	REG	AUM + Holiday + Weather
Multiple Regression	**3311.61**	**3225.04**
Regression Tree	6247.51	5694.43
Random Forests	6181.55	5975.22
Gradient Boosting	6102.49	5762.21
LSTM RNNs	7822.23	7266.45
GRUs RNNs	6578.34	6388.56