**CIND 820 Final Report**

**Student Name:** Justin Xu

**Student number:** 500936632

**Supervisor name**: Ashok Bhowmick

Submission date: 2021/10/16

**Associated Links:**

GitHub Repo: <https://github.com/tunstall0729/CIND-820/blob/main/CIND%20820%20data%20des%20latest%20version.ipynb>

**Table of Contents**

[Abstract 2](#_Toc87102970)

[Introduction 4](#_Toc87102971)

[Literature Review 5](#_Toc87102972)

[Dataset 11](#_Toc87102973)

[Model Implementation and Evaluation: 22](#_Toc87102974)

[Methodology: 23](#_Toc87102975)

[References: 25](#_Toc87102976)

# Abstract

**Research theme and topics:**

When considering purchasing a car, price is always the concern for most people. This project aims to create a model that could be used to *automatically* predict the price of used cars on the market with certain information provided.

The research seeks to answer following questions such as, how prices vary with different features of the car? which variables are significant in predicting the price and which ones are not? Does brand influence the price? The objective is to choose the appropriate parameters and then determine a model which could predict the price as accurate as possible.

The research theme chosen here is predictive analytics using techniques such as regression and classification. The goal is to apply the data analytics methods and knowledge I learned through previous course in a project setting and try to solve a real-world problem.

**About the dataset:**

The ‘Used cars listings for US & Canada’ datasets are sourced from Kaggle website([Used cars listings for US & Canada | Kaggle](https://www.kaggle.com/rupeshraundal/marketcheck-automotive-data-us-canada)). These two files contain used vehicle inventory data from different dealers across the US and Canada. The two dataset have over 7 million records in total and 21 attributes. Data types include Strings (Object as in Python), Integer and Float. Each record gives you basic information such as year, brand, and miles. It also has equipment breakdowns such as fuel type, engine size and transmission. Missing values were found after a preliminary analysis was performed; thus, data cleaning is required for further analysis.

**Proposed techniques and tools:**

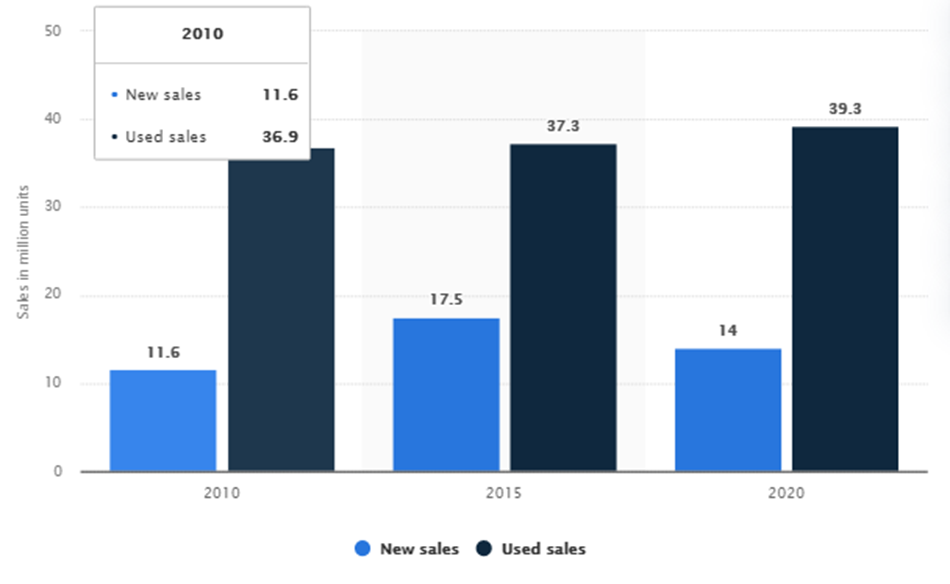
Python Jupyter Notebook will be used in this project for coding.

Exploratory data analysis will be performed to explore the dataset using data visualization tools like histogram, box plot and scatter plot. Data cleaning process will also be implemented to make sure data is of good quality for analysis. This process may involve locating missing values, outliers, and inconsistent data, and then correcting, modifying, or removing these records based on research purposes. The correlation between variables will then be identified in order to select the correct numbers of features for further analysis.

To achieve the research objectives, there is a need to study the relationship between outcome (price) with other independent variables(features). Then the dataset will be split into training data set and test data set. The basic idea here is to infer a function from labeled training data consisting of a set of training examples to map new examples, thus, the task is a supervised learning [1]. Techniques such as Linear regression, Decision tress and K-nearest neighbor algorithm will be used for analysis. Statistical measurement such as confusion matrix (error matrix) will be calculated to compare the results from different algorithms. The final step here is to determine the most appropriate model that can predict the price for new records as accurate as possible.

# Introduction

The used car market is an ever-growing industry. According to a report provided by Eric Rosenbaum, CNBC [1], the US used-car market is more than twice the size of the new-car segment and is outpacing it in growth. The used car market in the US is estimated at 41 million units annually. It is the second highest-priced asset consumers purchase; thus, it has a great importance for the economy. Furthermore, Covid-19 has led to an increase in used car sales as people avoid mass transportation and are more sensitive to auto cost in the recession. The statistics provided by statista [2] also illustrates (Fig 1) that the sales of used cars have outperformed new cars by a large margin throughout 2010 to 2020. It also states that due to high demand but short supply of used cars, prices will ~~also~~ keep rising. This adds additional significance to the problem of used car price prediction.



**Fig 1. New and used light vehicle sales in the United States from 2010 to 2020**

Price has always been one of the most critical factors when considering buying or selling a car for most people. The current second-hand car market has created opportunities and challenges for both buyers and sellers. Buyers prefer to buy used cars because of their relative affordable prices compared to new cars, but they are also concerned with car dealers taking advantages of them by listing inflated prices given the current high demand and short supply scenario. On the other hand, sellers need to put forward a reasonable price, so they will be able to sell their cars and make a profit at the same time. Therefore, one can see that estimating the price of used cars has a very high commercial importance. In this regard, a good predictive model may assist people in budgeting by having an accurate estimate of the would-be price of a vehicle with respect to various features in it.

Although car price prediction is an interesting and popular problem, it is actually not a simple task as it seems superficially. Good domain knowledge is required to understand the pricing dynamics because various features influence the accuracy in prediction of the car price. The major factors are normally brand, car model, the age of the car, mileage (the number of kilometers it has run), and horsepower. The fuel type used in the car will also affect the price due to rising fuel prices. Other features like transmission, color, the drive type whether it is 4wd or fwd, the number of cylinders and car type (e.g., SUV or sedan) will also influence the car price. The list of features is not limited, one can always add new features into the prediction.

# Literature Review

Several related works have been done previously on the subject of used car price prediction using different methodology and approach.

As a first case in point, Enis Gegic et al [3] first applied three algorithms separately on their dataset to predict car prices: Random Forest, Support Vector Machine and Artificial Neural Network. Then they applied an ensemble method that combines these three algorithms together on the same dataset again. The respective performances were then compared to find one that best suits the available dataset. The results indicate that the ensemble model has a much better accuracy compared to single machine learning algorithm. The dataset was collected from a web portal including used cars data in Herzegovina and Bosnia. The limitation of this research is that only 1105 records were included in the dataset and the computational resources required is huge.

Ashish Chandak et al [4] applied K Nearest Neighbour (KNN) and Regression Trees in their research for car price prediction. The root means square error for KNN with K=7 is 5581.96 and for Decision Trees is 4961.64. The issue with their research is that a few important features were not included in the dataset. Detailed information about the dataset were also missing.

In Sameerchand Pudaruth’s study [5], four different techniques: multiple linear regression analysis, k-nearest neighbors, naïve bayes and decision trees have been used to make the prediction for used cars in Mauritius. The predictions are then evaluated and compared in order to find those which provide the best outcome. The dataset was collected manually from local newspapers in period less than one month. After further pruning, they only kept the three most popular makes in Mauritius. The performance of each model was not optimal with overall accuracy below than 70%. The drawback of his research was again the low number of records and features that have been used. Pudaruth also concluded that Naïve Bayes and Decision Tree are unable to handle output classes with numeric values very well.

K.Samruddhi et al [6] conducted their car price prediction by using KNN (K Nearest Neighbor) regression algorithm. The dataset used for the model was collected from the Kaggle website. The data was trained and tested using different ratios by the model with different K values. The same model is then cross-validated for inspecting overfitting of the model using 5-Fold and 10-Fold method. The highest accuracy rate is 82% when K = 4, however, the number of records of the dataset was not mentioned in the paper.

In their 2019 study, Pattabiraman Venkatasubbu et al [7] proposed using Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees to develop a model which will be able to predict the price of a used car. The accuracy of these models was then compared to determine the optimal one. The data set used for prediction models was collected from the 2005 Central Edition of the Kelly Blue Book and has 804 records of 2005 GM cars. The prediction error rate of all the models was under 5%. But, on further analysis, the mean error of the regression tree model was found to be larger than the mean error rate of the multiple regression and lasso regression models. The ANOVA test has also confirmed that even though for some seeds the regression tree has better accuracy, overall, its error rates are higher compared to the rest. However, the limitation with this research was again the small number of records collected. Also, only GM cars were included which certainly causes a bias since car brands induct a huge impact on the price.

After studying the related work shown above, it is the author’s observation that single machine learning algorithm approach seems not able to generate remarkable results with small number of instances. It is interesting to find out if the result is mostly influenced by number of records or due to the natural of simple algorithms.

This paper set out to study, given sufficient data samples and features, if advanced machine learning algorithms are more powerful compared to simple algorithms like KNN and linear regression in predicting used car prices.

Compared to above mentioned studies, the dataset collected from Kaggle website for this project provides adequate numbers of records and features for study. The variety of records will help to reduce the bias factor. The dataset will first be trained and tested with linear regression. Then ensemble learning techniques like random forests and XG boost will be applied on the dataset. Afterwards, the results of each model will be compared to determine the most suitable one for prediction. Finally, the same method will be used on a new dataset for a different city. The results will then be compared to see if location will make any difference in price prediction.

The basic idea of ensemble methods is that a “weak” learner can be improved significantly if given an opportunity to operate “by committee” [8]. The two major techniques under ensemble methods are Bagging and Boosting.

Bagging stands for “bootstrap aggregation”. It assigns cases to categories by majority vote over a set of bootstrapped classification trees so that it can adjust the overfitting problems.

Random Forests is an extension of Bagging technique. It is a combination of decision tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [9]. Significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. In Breiman’s study in 2001 [9], Random Forest was able to produce significantly lower error rates, especially on larger datasets, which suggests that injecting the right kind of randomness can improve the results. However, it is worth mentioning that, just like bagging, there is no longer a single tree structure to interpret. Therefore, it is hard to learn which predictors are driving the outcome.

On the other hand, Boosting is a forward stagewise additive model [8], while decision tree works with smaller and smaller partitions of the dataset at each stage, boosting uses the entire data set at each stage. It “combines the outputs from many ‘weak’ classifiers to produce a powerful committee”.

Ensemble methods have already been used extensively on different research topics regarding prediction problems.

In 2021, Dita et al [10] compared the performance of linear regression and Random Forest in predicting sneaker resale prices using sales history data gathered from StockX. The conclusion is that Random Forest model with 10-fold ~~of~~ cross-validation performs better compared to the Linear Regression model in predicting the price. It is also stated that because of Random Forest’s ability to handle outliers which cannot be found in Linear Regression, it makes the case more suitable to work with complex data just like the one used in this study.

Samir et al [11] built an energy consumption baseline model based on a gradient boosting machine to predict commercial building electricity consumption. Hyper-parameters tuning techniques were used to refine the model. The results showed that using the gradient boosting machine model improved the R-squared prediction accuracy and the CV(RMSE) for more than 80 percent of the cases when compared to an industry best practice model that is based on piecewise linear regression, and to a random forest algorithm as well.

One comprehensive research on this topic was performed by Stefan et al [12] in 2017. The study has evaluated the comparative performances of 19 regression methods for forecasting resale price in the used car industry. These regression methods are grouped into individual and ensemble methods. The sales dataset contains six different car models in the second-hand market was provided by one leading German car manufacturer.

The experiment was split into two settings: high-dimensionality and low-dimensionality. Different set of variables were chosen for each setting. The results show that ensemble methods are the overall winners of the comparison across several experimental conditions. These methods predict resale prices significantly more accurate than individual prediction methods due to their ability to capture the nonlinearity in a data-driven manner. The analysis also provides evidence that random forest regression is particularly effective for price prediction. The fact that many methods predict resale prices fairly accurately, suggests that high dimensionality was not a major obstacle. Sample size in fact plays a more important role when estimating complex, nonlinear relationships between covariates.

This project will make its contribution by performing a systematic comparison of several widely used and diverse supervised machine learning methods (Individual vs Ensemble) to provide evidence on how each method differ in their predicting accuracy, or which method is most effective. The overall approach will be somewhat similar to the one used by Stefan et al but using a more comprehensive and recent dataset. Compared to their research, this project will include other brands of cars in study and see if the conclusions will still hold true.

# Dataset

1. **Description**

The ‘Used cars listings for US & Canada’ datasets are sourced from Kaggle website([Used cars listings for US & Canada | Kaggle](https://www.kaggle.com/rupeshraundal/marketcheck-automotive-data-us-canada)). The dataset is licensed under Public Domain, which means one can copy, modify, distribute, and perform the work. The datasets were first available to the public in 2021 and is scraped on a monthly basis. The current version of this dataset is 5th. The two files are downloadable in SCV format.

The Canada dataset has 393,603 records while the US dataset has 7,104,304 records. Both datasets have 21 attributes. Only records under the city of Boston and Toronto will be selected in in this project for comparision. Attributes that are not of interest to the study of this project like demographic will also be dropped. Dropped attributes are listed as below:



The dataset that will be used after trim has 13 attributes. Data types include Strings (Object as in Python) and Float. Each record gives you basic information such as year, make, and miles. It also has equipment breakdowns such as fuel type, engine size and transmission.

A picture containing text, receipt

Description automatically generated

**Data attributes and definitions:**

|  |  |
| --- | --- |
| Price | Price of the car. |
| Year | Model year of the car |
| Make | Manufacturer/Brand of the car |
| Model | Model of the car, e.g., Civic, NSX… |
| Trim | Trim levels are used by manufacturers to identify a vehicle's level of equipment or special features |
| Fuel\_type | Fuel type of the car. |
| Miles | The mileage. It measures the distance traveled by a vehicle. |
| Body\_type | Body type of the car |
| Transmission | Transmission of the car, Manual, Automatic… |
| Drivetrain | Drive type of the car, e.g., rwd, 4wd… |
| Vehicle\_type | Vehicle type of the car, eg, Car, Truck |
| Engine\_size | Engine size of the car |
| Engine\_block | The type of the engine block |

**Dataset Source Literature:**

The dataset provides multi years of inventory across the US and Canada. The data is crawled and aggregated from over 65k dealer websites to deliver the most comprehensive and up-to-date depictions of market activity.

Individual listing records show year, make, model and trim, with VIN-level histories, showing the most recent time the car showed up online back to the earliest, with every change that occurred over that time.

Equipment breakdowns give fuel type, engine size, transmission, color, driveline and body style.

**Sample data:**

**Table, Excel

Description automatically generated**

Missing values were detected among different attributes. These missing values will be analyzed and handled in the next stage.

A picture containing text

Description automatically generated

1. **Exploratory Analysis**

Exploratory Data Analysis is used to understand the dataset, to gather insights, perform initial investigation and discover anomalies and patterns by using different visualization techniques.

The price attribute is the key attribute here since it is the only dependent variable for this research. A boxplot is created, and it illustrates that prices are right (positive) skewed which means they are not balanced.

Chart, box and whisker chart

Description automatically generated

After including outliers in the boxplot, one can see that there are quite a few records with extreme high prices which explains the skewness of the distribution.

Chart, box and whisker chart

Description automatically generated

Another boxplot comparing price against manufactures was then performed to try to locate these outliers. Based on the graph, one can see that manufactures like Ferrari and Rolls-Royce have a much high median price and max price due to the nature of their brand. These outliers will also need to be studied and handled properly.

Chart, histogram

Description automatically generated

Some other initial data exploratory analysis as follow:

Top 10 count by manufacturer and model:

Chart, funnel chart

Description automatically generated

Top 5 body type, drivetrain, and transmission:

Chart, bar chart

Description automatically generated

Number count by year:

Chart, bar chart

Description automatically generated

1. **Data cleaning and preprocessing:**

Data cleaning and preprocessing is the first and most important step in the process of developing our predictive models. Generally speaking, real world data may contain noises, missing values and outliers which cannot be implemented directly in algorithms.

**a). Checking null values:** Improper handling of missing values will negatively impact the performance of the models. There are missing values in both datasets. Various techniques such as imputation methods are used to replace null values with median or most frequently values. Figure shows the missing values in each attribute in Toronto dataset.

Text

Description automatically generated

**b). Detecting outliers:** The outliers are usually data points which deviates far from the majority. Outliers may occur in dataset due to human or machine errors while collecting the data. By using various visualization tools such as boxplot and scatterplot, we can identify the outliers in our datasets:

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Outliers can also affect the model performance. The outliers are removed in both datasets to improve the accuracy in predicting the prices of cars. Figure and Figure demonstrate outliers in miles and price.

**c). Inconsistent data:** Inconsistent data points were also discovered in both datasets as shown in Figure as an example. Some vehicle types were categorized incorrectly. Another example is that electric cars were entered with engine block and size information. These data points are corrected by using common sense and domain expertise.

Table

Description automatically generated with medium confidence

**d). Feature Scaling**: Features with large data range will dominate other features with smaller range and may hide valuable insights. To avoid such issue, all the numeric features in both datasets are scaled and normalized using min max scaler so they can contribute equally to the models.

A picture containing text

Description automatically generated

**e). Encoding Categorical Variables**: Categorical attributes can not be applied directly into algorithms. Attributes like model, drivetrain and body\_type need to be transformed first. Most of the categorical features in the dataset have more than two values. If we use LabelEncoder then these values will be treated as ordinal ones by the machine learning model which will induce bias. Therefore, one-hot encoder, also known as dummy encoding, would be a proper choice, however, the ‘model’ feature here has a high cardinality with more than 500 values, in order to avoid curse of dimensionality, we need to diminish the high cardinality. The approach used here is to apply rare encoder on model attribute first. All models that appear less than 20 times were labelled as 'Rare'. With the help of rare encoder, the cardinality decreased significantly. Then one-hot encoder was applied on all categorical attributes.

**f). Train and Test Splits**: Each dataset was split into train and test with 80% training and 20% testing as shown in fig:





# Model Implementation and Evaluation:

1. **Linear Regression**:

Linear regression predicts the value of dependent variable based on a given collection of independent features using linear predictor functions. The algorithm was applied on both Toronto and Boston datasets. R2 (root square) and RMSE (root mean square error) were then calculated to evaluate the results.

The Toronto dataset has a R2 with 82.62% and RMSE with 5576.62. While the Boston dataset has a R2 with 87.36% and RMSE with 5122.60. To test whether the model is overfitting, the algorithm was applied again on the training dataset and the scores were calculated as well. With similar R2 and RMSE score, it seems that overfitting is not a problem.

|  |  |  |
| --- | --- | --- |
| Linear Regression | R2 | RMSE |
| Toronto (Test) | 82.62% | 5576.62 |
| Toronto (Train) | 82.18% | 5725.81 |
| Boston (Test) | 87.36% | 5122.60 |
| Boston (Train) | 86.40% | 5304.94 |

# Methodology:

Data Exploratory

Analysis

* Missing Values
* Outliers
* Feature Engineering

Data Cleaning/Processing

Split dataset into Training and Testing

Apply Individual Method:

Linear Regression

Apply Ensemble Method:

Random Forest and Gradient Boosting

Testing and Performance Evaluation with metrics

# References:

1. Mehryar Mohri, Afshin Rostamizadeh, Ameel Talwalkar. *Foundations of Machine Learning*, The MIT Press,2012, ISBN 9780262018258
2. Eric Rosenbaum. *The used car boom is one of the hottest, and trickiest, coronavirus markets for consumers.* CNBC EVOLVE, 2020. Available at: [Used car boom is one of hottest coronavirus markets for consumers (cnbc.com)](https://www.cnbc.com/2020/10/15/used-car-boom-is-one-of-hottest-coronavirus-markets-for-consumers.html)
3. Mathilde Carlier. *U.S. new and used car sales 2010-2020*. Statista, 2021. Available at: [• U.S. new and used car sales 2010-2020 | Statista](https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/)
4. Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric. *Car Price Prediction using Machine Learning Techniques*. TEM Journal, Volume 8, Issue 1, 2019, page 113-118. Retrieved from: [CEEOL - Article Detail](https://www.ceeol.com/search/article-detail?id=746689)
5. Ashish Chandak, Prajwal Ganorkar, Shyam Sharma, Ayushi Bgmar, Soumya Tiwari. Car Price Prediction Using Machine Learning. International Journal of Computer Sciences and Engineering (JCSE), volume 7, Issue 5, 2019. Retrieved from: <https://www.ijcseonline.org/ijcse_search.php?txtsearch=Car+Price+Prediction+Using+Machine+Learning&submit=Search>
6. Sammerchand Pudaruth. *Predicting the Price of Used Cars using Machine Learning Techniques*. International Journal of Information & Computation Technology, Volume 4, Number 7, 2014, page 753-764. Retrieved from: [2014\_Predicting\_the\_Price\_of\_Used\_Cars\_using\_Machine\_Learning\_Techniques-with-cover-page-v2.pdf (d1wqtxts1xzle7.cloudfront.net)](https://d1wqtxts1xzle7.cloudfront.net/54261672/2014_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Techniques-with-cover-page-v2.pdf?Expires=1634176644&Signature=UhU8vGraK0VqPMA9PZ6dC5FdZoYEq3iPKuf07In6P9qA4wFk6oTlKg1ZwE1wzDqi2hlTXSPca2U9JKo~UHEDJB-uK4yjNLJDaL8ZFjYep20OM2pS1~nK-3slba6VUhKrQdefLNYWsPC73pWNvLj8gMD9-OfCgBXt5PuN7f8lWjXPG0e9cOd8EEV0YCcTNEkRH4fXc-xNAYrtBQCe~O~A6Xxea7FVHPJlrX3Irs21wU8e~AU6ZS1nQ9ogIeXZ2x~pWkob-En7dOdYNbadIxIgF4tNitYX3JfL-hrPkWyTjk8bM3Q2oKDzPsioeQ0RmXQ7PDG6~HlPY54czLK3hZ92Kg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)
7. K.Samruddhi, Dr R.Ashok Kumar. *Used Car Price Prediction using K-Nearest Neighbor Based Model*. International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE), Volumn4, Issue 2, 2020. Retrieved from: [V4-Issue-2-629-632.pdf (ijirase.com)](https://ijirase.com/assets/paper/issue_1/volume_4/V4-Issue-2-629-632.pdf)
8. Pattabiraman Venkatasubbu, Mukkesh Ganesh. *Used Cars price Prediction using Supervised Learning Techniques*. International Journal of Engineering and Advanced Technology (IJEAT), Volume 9, Issue 1S3, 2019. Retrieved from: <https://www.ijeat.org/wp-content/uploads/papers/v9i1s3/A10421291S319.pdf>
9. Richard A.Berk. *An Introduction to Ensemble Methods for Data Analysis*. SAGE Journals, Volume 34, Issue 3, 2006, page: 263-295. Retrieved from: <https://journals-sagepub-com.ezproxy.lib.ryerson.ca/doi/10.1177/0049124105283119>
10. Leo Breiman. *Random Forests*. Machine Learning, Volume 45, Issue 1, 2001, page: 5-32. Retrieved from: <https://link-springer-com.ezproxy.lib.ryerson.ca/article/10.1023%2FA%3A1010933404324>
11. Dita Raditya, Nicholas Erlin P, Ferarida Amanda S, Novita Hanafiah. *Predicting Sneaker Resale Prices using Machine Learning.* Procedia Computer Science, Volume 179, 2021, page: 533-540. Retrieved from: [Elsevier Enhanced Reader](https://reader.elsevier.com/reader/sd/pii/S1877050921000430?token=5845AF0561EBF5B1A85942E3C6332BB752A226ED9698EAC7C3F5BA12D0061ED63631C563A495B55D59F0B9441D290D49&originRegion=us-east-1&originCreation=20211010194339)
12. Samir Touzani, Jessica Granderson, Samuel Fernandes. *Gradient boosting machine for modeling the energy consumption of commercial buildings*. Energy and Buildings, Volume 158, 2018, Pages 1533-1543. Retrieved from: [Gradient boosting machine for modeling the energy consumption of commercial buildings (sciencedirectassets.com)](https://pdf.sciencedirectassets.com/271089/1-s2.0-S0378778817X00200/1-s2.0-S0378778817320844/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEHkaCXVzLWVhc3QtMSJGMEQCIH%2BkJ0Tmox2UHB1u9gFjf2rvCJXlg7tViLn9Y9HXUhY%2BAiBe6GbCmradxst1SLVJnHtmg47niLMiT7FcczWI1yq9QSqDBAjy%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F8BEAQaDDA1OTAwMzU0Njg2NSIMellxMTQNJsd9DQbaKtcDpBKu08etBLrtBLxrYvN5hMRSzdEYRBH69tYA0ovDsTyHBwHcrWRxPKkEvTDHk221v2KxQVv92flFSqZcOrwreHNhjIyXp2RbMx6%2FVOhpdBDTK9m5pzL0UA0HJHjQoKk9ntcZ8UB90dtHB6IqzwLu7Dbj2g%2F8x4DP3GILmMzwGZzUfywp3o%2Bwt5fZHYO7oPRLbQqSJupbyX4yaH2TcchYbm7sj%2BjDgYdE9Y8b%2FIdC3s%2BCvuuaaGP07ciyPEiupMCLY38yl3KdcKzy%2FGHS73vAIwyHG1F4bEp0GAvD8Stab9MkZKujURdhrn35AnaBEWsmF7AZrUB9%2FuPYbQtmWh9hdrfQuAoVb7zyDRwVirrcXyQV3GW9dBqBNmf0DcNehfHtX6htVqdazOP7SgropYrSuP7f3UOqZx5lozAmjFkTfZqcg9asUlW6Ip26mbUnOjc1lfzd4%2Bqd727UDj89BfVy9oRsdClfHHsjxF9zvoipCnz29xJC6yvrLAnQBq7IbV01%2FLksG9vAfr8Lt0aErQow5ESpuMP%2BRfY85zOlSZzNhjxViqqgu5NgFaSEvUjiyGVFUH7tPnivvKrolXXWUhHIvUb60Dxlv4W2phntuzA6YnXOZLyQLnfEMM%2FZkYsGOqYBegOxAXtGzqeYonRh%2FptUq%2FI7cZ%2BORzsS2mQXlUgVqRwImT6kWsxnEZDglsSwcgnkK4jeN5IumB1D8uPpOHdTpamyF5SFcHXHj2f7uFxV96bboSqpHIcRY4C6gyefT8qcpAHqj5yDVgGmrZRpX%2FjlyJa9DrJrLccMInbZ0g6%2Bp4GdlI8E1IN3fHOofCNbMkrYYp%2BIVlueHZ%2BdewTZi%2BlagCpbdYC0tg%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20211011T175552Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTY3TVJVCNA%2F20211011%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=b7106c8b1f0866035738f94ed728ddc3a1ad9eff0d6be05eef99ba313445ebec&hash=c674e3ed4e9870c12d1c58e1359ddbaa36a95532e21cd9a0d58f68ceb28b3aa7&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S0378778817320844&tid=spdf-8dc60427-bfe4-41f9-a969-8c9631bb489d&sid=275b36b847ec1044f3997e3-8b5781c227a7gxrqa&type=client)
13. Stefan Lessmann, Stefan Voß. *Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy*. International Journal of Forecasting, Volume 33, Issue 4, 2017, page: 864-877. Retrieved from:https://link-springer com.ezproxy.lib.ryerson.ca/article/10.1023%2FA%3A1010933404324