Used cars price prediction, Prediction of New York taxi trip cost, and the salary of NYC employee

Rupesh Sai Baba Chintakayala
M.Sc in Data Analytics
National College of Ireland
Dublin, Ireland
x21104638@student.ncirl.ie

***Abstract*—This report provides a fundamental analysis of the cost of commuting for an average income person in New York City for the purpose of the Data mining and Machine Learning module semester project at the National College of Ireland.**

When the COVID-19 shutdown happened across the world, many industries lost their position in the market and one of them was the taxi industry as all the commuters were working from home, no tourists and other businesses were also shut down. When cities started recovering from the pandemic, the immense rise in demand for taxis to commute for everyday work has created a huge gap between the demand and supply of the taxis and passengers. The taxi industry was not ready to meet the demand of the commuters for daily life, in turn rising the base fares for taxi rides making the average-income population hard to afford. This raises the question of either buying a used car or using a taxi to commute. The objective of this project is to improve the accuracy of prediction of the taxi trip cost, price of a used car, and salary of different individuals from one of the busiest cities across the world i.e., New York using the KDD methodology by making use of Python, PostgreSQL and draw essential insights that could help to acknowledge the problem and get to conclusions to address the problem.

Keywords—Knowledge Discovery in Databases (KDD), CSV, Python, Taxi duration, used cars, payroll.

I. Introduction

There are so many different options to commute to work these days, you can drive on your own, hire a taxi, or take a bus/ train (public transport), but it depends on the individual and the time they can spend commuting to work on a given day. Out of all these options, people tend to choose either driving on their own because they can spend listening to podcasts/ music and have control overtime or taking a taxi to beat the rush of crowd and traffic with public transport.

Figure 1: Transportation in New York

Owning and maintaining a vehicle in one of the busiest cities in the world is not that simple, and likewise affording a taxi is also becoming next to impossible due to high demand and very high prices these days. Since COVID-19 has displaced all the workforce to work from home and the people who live on fieldwork had to shut down their businesses and work. The idea of this project is to analyze the data from taxi trips, used cars price, and salaries of different individuals in New York City, and apply various machine learning algorithms to know which performs better to predict future values to help us understand the consumer behavior on the choice of transportation for the commute. To support the assumption of commuters of choice of transport is based on the income they are getting and the fares of public transport, taxi fare, and own transportation, this project helps us understand a few of the components of the above problem to get started.

The main objective of this project is to improve the accuracy of the prediction of used cars prices, the price of a taxi trip, and the salary of New York City employees using different machine learning algorithms and compare the behaviors of these algorithms on different datasets.

II. Related work

In the paper [1], the author predicts the price of used cars by using four different machine learning algorithms - k-nearest neighbors, Naive Bayes, Multiple Linear regression, and decision tree algorithms. The historical data collected from daily newspapers in Mauritius is used by the author for this research. The listed machine learning algorithms were applied to the data and the results obtained didn't have good accuracy. In this research paper, the author was able to achieve an accuracy of 60-70% and failed to achieve better accuracy. This could be achieved using Random Forest, Decision Tree Regression, Lasso Regression, SVM (Support Vector Machines), etc.

Nabarun Pal et al. (2018) [2], have used the random forest algorithm to predict the price of used cars and were able to achieve 95.82% training accuracy and 83.63% accuracy in testing accuracy. The authors have created a number of decision trees based on the Grid search algorithm to find the optimum number of trees to train data and were able to achieve a better accuracy score compared to previous studies/ research on the same topic and for the same data. This accuracy could further be improved by considering more features from the dataset as in this study, they were confined to a few features such as kilometers, vehicle type, brand, and price. Some other features which can be considered are fuel type, the volume of cylinders, safety index, etc.

In the paper [3], the authors have used the Adaptive neuro-fuzzy inference system (ANFIS) and Artificial Neural Network (ANN) with backpropagation (BP) because of their adaptive learning capability to predict the price of used cars and compared both price forecasts. The system they proposed has three different stages: data acquisition, price forecasting algorithm, and performance analysis. Finally, the authors were able to achieve the convergence values of 0.83477 and 0.6531 from the training process of BP and ANFIS respectively. Since results are based on limited data, limited features, and high computation complexity, this affects the implementation of this in real-time scenarios as the knowledge from the rest of the features might be lost during the analysis. As the number of features increases, the training time increases, and also it would be computationally very expensive.

In the paper [4], the authors proposed an automated technique to predict the used car price using Random Forest, XGBoost, and LightGBM. They also used entity embedding while using the above-mentioned machine learning algorithms and observed the difference between the values of R-squared with and without entity embedding. The authors have tested the proposed model on two different datasets, one is the used cars dataset from Kaggle and another is the Vietnamese dataset (collected from newspaper pages). The accuracy was better with entity embedding rather than using a single algorithm on the dataset to predict the price. However, the accuracy (0.8634) obtained was not that good and it could have been better to use a decision tree along with Random forest and try entity embedding on top it might have made a difference in inaccuracy.

In the paper [5], the authors have used model selection, lasso regression, and random forest to predict the duration and fare of the taxi ride. To predict the fare, they used the data which was available before the ride starts such as - pickup coordinates, drop-off coordinates, trip distance, number of passengers, type of rate code, and start time. The linear regression model was failing to improve prediction after a certain point and the performance was limited due to uncertain traffic patterns. Out of all the models built using machine learning algorithms used in the study, random forest outperforms all other models as it was able to overcome the challenges such as variation of traffic (non-linear) and uncertainty in location. However, this paper does not provide any statistical figures regarding the model's accuracy and performance. Apart from this, the paper only emphasizes the use of different features and how they are used in fare prediction.

Balika et al. (2021) [6], have proposed a model to overcome the problem of fare amount prediction for taxi commutes. They used two different formulas to calculate the distance between two points to calculate trip distance. They are the Euclidean distance and the Haversine Formula. They designed a model with a few different modules which consist of data evaluation, pre-processing, and prediction. They didn't implement this model on any dataset but provided a different view on this problem statement on how this could be done from a different perspective. They mentioned key feature extractions which consist of Partial feature extraction, Temporal feature extraction, and Dynamic feature extraction, these might help us boost the accuracy and predict the price precisely. However, the other factor which might affect this method is peak traffic hours, and many different routes available between two given points. These can be handled only with vast data related to traffic, routing, and weather, which is also one of the factors to be considered in fare prediction. The model mentioned in this paper utilizes taxi-request indexing for improving the matching of taxi and passenger. As per the evaluation results mentioned in the paper, the accuracy of the model showed good results and the model was quick in processing the requests even at the milliseconds level. However, this paper does not provide any statistical figures regarding the model's accuracy and performance. Apart from this, the paper only emphasizes the use of AI, and no study was provided regarding the work of conventional machine learning techniques.

In this paper [7], the authors have estimated the taxi fare by building a model with dynamic traffic conditions and various routes. They built three different models, each model dealt with different conditions and was used to build the next model by using the previous one. In the first stage, they built a model to analyze taxi drivers' behavior in routing. Using this model, they analyzed the income patterns of the taxi driver. Next, they built a model to understand traffic patterns. Using these two models, they employed a tucker model to estimate fare for any given route. They also used a smoothening method for the final model to achieve better accuracy. This study helps us understand the problem from a different perspective and consider all the factors affecting the problem, but it fails to include other machine learning models such as Random Forest, SVM, Linear regression, and AdaBoost for a more comparative study.

In the paper [8], the authors have developed an application and used two methods to improve the accuracy of estimating taxi fares. They used adaptive calibration to improve the accuracy of estimating taxi fares when running the prediction program. They used an improved distance accuracy method to reduce the error between the cumulative distance in the program. Their application also uses messages, social networks, and cloud storage. The major functions of this application are quotes as fare estimation, real-time route planning, and a security tracking system. Using this application in real-time they were able to achieve an error rate of 9% and 6% during peak and off-peak hours respectively. The main drawback of this approach is that the user had to bear additional costs linked to messaging services and service providers and the user had to run this application on a mobile phone for this estimation to work which is a new behavior to the user to learn.

In the paper [9], Anzhelika Antipova has analysed how low-income workers suffer a huge commuting cost burden compared to normal commuters in the context of economic opportunities. In the study paper [4], the authors evaluate whether a low-income working person suffers a high cost of commute in contrast to a typical commuting person with reference to the decreasing opportunity. In addition to this, the study also offers comprehensive insights on the methods to change the location of work and population in a metropolitan area which may affect the commute distance of low-income traveling people. The statistical figures reveal that low-income workers commute shorter distances to their workplaces. This study however is limited to only one particular city and hence a general scenario of the problem cannot be grasped.

In the paper [10], the authors employed the PCA and Support Vector Machine method to predict the income of a person. The performance of the SVM model was then compared against six distinct subsets of the dataset. Accuracy of 84% was achieved when compared to the test data. This study however fails to include other machine learning models such as Random Forest and Linear Regression for a more comparative study.

In this paper [11], Swapnajit Chakraborti has proposed five distinct machine learning models to predict the salary of an individual. Five Machine Learning algorithms namely - were used for modeling. The study also provides a comprehensive and comparative analysis of the performances of the other algorithms. From the results, the decision tree model performed slightly better in contrast to other models and predicted 86.21% of the instances correctly.

In [12], The study explored Zipcar's vehicle utilization pattern in NYC and analyzed the usage of carsharing based on different socio-demographics in neighborhoods. Features such as number of total vehicles, number of parking lots, accessibility and economic background of the neighborhoods, income class, etc. were analyzed. It was evident that the subsidized pricing model based on neighborhood locations and an increasing number of service vehicles made the model more feasible. Even though carsharing can be affordable at times, the availability of the vehicles at a certain time is still an issue. The study made a comprehensive review of the carsharing model, however, the data used was inconsistent and unreliable. An expansive data collection using specific feature selection that includes furthermore predictors such as the reason for renting carshare, hourly rates, and demographic data - family income, employment status, race, etc. could have elaborated on the feasibility of the study.

III. Methodology

Knowledge Discovery in Databases (KDD) is the process of discovering unknown patterns and useful knowledge from data. This process involves the repeated applications of a few different steps which will be discussed now.

A. *Domain understanding and KDD goals*

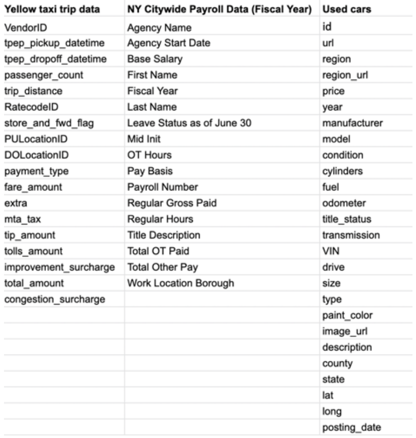
The domain of this project is transportation (to be specific, commute to work for an employee) and it aims to predict the price of used cars, prediction of taxi fares in New York City, and prediction of salary of New York employees. The goal is to understand the behavior or preference of an individual on the choice of commute to work or for any other daily work. The selection of data should also be done in such a way that it supports different aspects of this study and help us understand the major factors affecting the choice of transportation for the commute.

B. *Data selection and description*

1. Used cars dataset: The used cars data is taken from the Kaggle website. This data set is created by Craigslist by web scrapping every few months. Craigslist is the world's largest collection of used automobiles for sale. This data contains 400k instances and 26 features ranging from the post date, color, and model to the price of the car. This dataset can be used to analyze the used cars data and predict the price of a used car.

2. NY yellow taxi trip data: The yellow taxi trip data is taken from New York City Taxi & Limousine Commission (TLC). The yellow taxi trip record data contains 3.2 million instances and 18 features including pickup and drop-off dates and times, pickup and drop-off location, trip distance, passenger count, payment type, fare, and other factors affecting total fare amounts such as tip, tolls, congestion surcharge, and improvement surcharge. This data can be used to analyze the taxi data and predict the fare of a taxi trip and the duration.

3. NYC payroll data: This dataset is taken from the Kaggle website. This dataset contains 3.9 million instances and 17 features like salary, regular hours, work location, leave status, base salary, work location, and name of every NYC employee. This data can be used to analyze and predict the salary of different individuals of different categories in the cities.



Yellow taxi trip data	NY Citywide Payroll Data (Fiscal Year)	Used cars
VendorID	Agency Name	id
tpep_pickup_datetime	Agency Start Date	url
tpep_dropoff_datetime	Base Salary	region
passenger_count	First Name	region_url
trip_distance	Fiscal Year	price
RatecodeID	Last Name	year
store_and_fwd_flag	Leave Status as of June 30	manufacturer
PULocationID	Mid Init	model
DOLocationID	OT Hours	condition
payment_type	Pay Basis	cylinders
fare_amount	Payroll Number	fuel
extra	Regular Gross Paid	odometer
mta_tax	Regular Hours	title_status
tip_amount	Title Description	transmission
tolls_amount	Total OT Paid	VIN
improvement_surcharge	Total Other Pay	drive
total_amount	Work Location Borough	size
congestion_surcharge		type
		paint_color
		image_url
		description
		county
		state
		lat
		long
		posting_date

Figure 2: Overview of the features from all datasets

C. *Data preprocessing and cleaning*

Pre-processing and cleaning should be performed in order to increase data reliability. It involves the steps such as handling the missing values, noise removal, outlier treatment, etc.

1. Used cars dataset: Initially the data is read from the source (CSV file) as a data frame. The features which have no use in predicting the price are dropped such as *id, url, region_url, title_status, VIN, image_url, and description*. Next, the number of null values is counted for all the features and since the size of the dataset is 400k, the features with null values of more than 300k are dropped such as *county* and *size*.

2. NY yellow taxi trip data: The data is loaded from the CSV file to a data frame and the null values are counted for all the features. The features which are having the null values are essential for the prediction of taxi trip prices. So the null values are filled with the mean of that respective feature and one of the features which is a categorical variable has null values of more than 100k, those instances are dropped from the dataset leaving the data size with 3.1 million instances.

3. NYC payroll data: After reading the data from the CSV file, the features such as *Last Name, First Name, and mid-Init* were dropped as they don't have a meaning in themselves to predict the salary of an individual. Since the data size is very large, all the fields with null values were dropped to reduce the size of the data and the computation power is also less to process very large datasets.

D. Data Transformation

To apply the machine learning algorithms, we must normalize the feature data for all the datasets. Most of the algorithms work well the numerical inputs rather than strings and objects. LabelEncoding is a simple function to use, that converts each value in a field/ column to a number. All the datatype of the features is converted to int to apply machine learning algorithms. The *value_counts* are taken for all the response variables and filtered based on the range with more values.

1. Used cars dataset: The response variable for the used cars dataset is price. So, the value_counts for the price variable gave the various range of values. So, to reduce the dimension of the dataset, we confined our price data from 1000 to 40000.

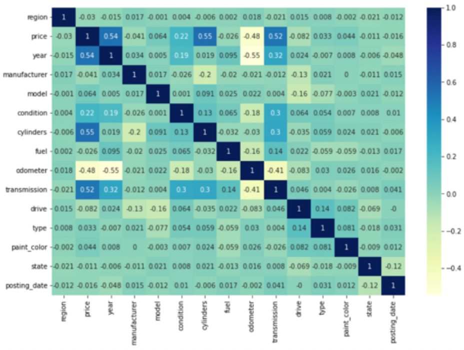


Figure 3: correlation matrix of features using heatmap in used cars dataset

2. NY yellow taxi trip data: The main variable for predicting the taxi price is the *total_amount* and the factors that are affecting the fare such as *fare_amount, tolls, tips, congestion surcharge, and extras*. So, all these variables should be greater than 0 and cannot be negative. This helps us remove the data which is invalid and reduces the dimension of the data.

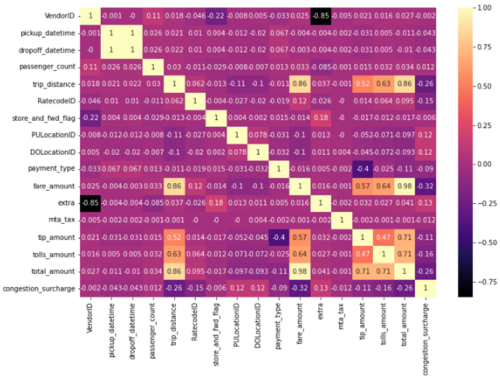


Figure 4: correlation matrix of features using heatmap in NY taxi dataset

3. NYC payroll data: The variables such as Regular gross paid, Regular hours, OT hours, Total OT paid, and Total other pay is crucial in deciding the salary of an employee, these variables cannot be negative. So filtered the data based on *value_count* for these variables and confined it to a range of 10000 to 80000 for regular gross paid and for OT, OT paid, and total other pay values should be greater than 100. The frequency of values below 100 for OT and for values outer the salary range are very less, so they are dropped to reduce the size of the data and to improve the accuracy.

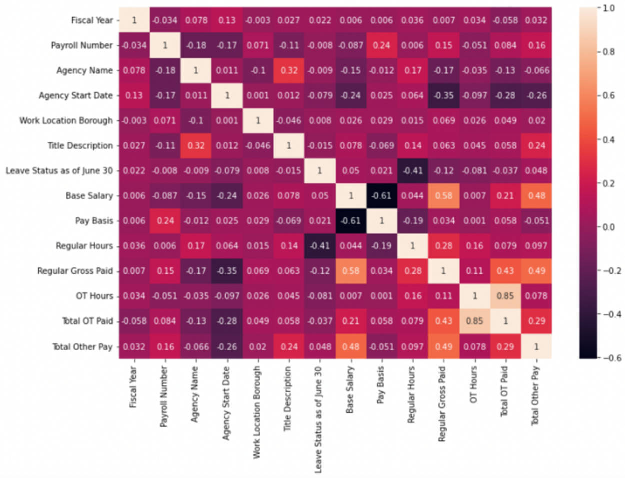


Figure 5: correlation matrix of features using heatmap in NY payroll dataset

E. Data Mining

As we got a clean and transformed dataset. We are ready to perform various machine learning algorithms with proper parameters to obtain better accuracy scores and fewer value in errors. The problem for this project is a regression since we are predicting the price of used cars, predicting taxi fares, and predicting the salary of NYC employees. So, we choose a few machine learning algorithms to implement on these datasets, they are Linear Regression, Random Forest, Ridge regressor, Bagging regressor, AdaBoost Regression, Stochastic Gradient Descent, and Voting regressor.

F. Data Evaluation and Interpretation

1. Used cars dataset: Out of all the models implemented on this dataset, Random Forest outperformed all other models with an r2 score of 98.69. Figure 3 shows a graph that is plotted with various values obtained by implementing different machine learning algorithms on the used cars dataset.

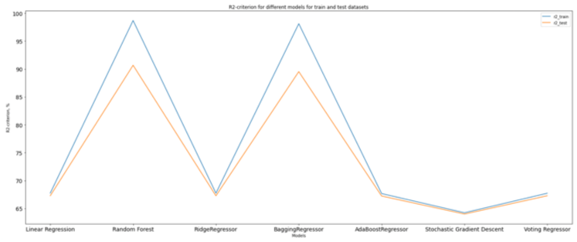


Figure 6: Used cars data - r2 graph for train and test for different models

2. NY yellow taxi trip data: All the algorithms performed well on NY taxi data with an accuracy of an r2 score of 99.78. From figure 4, we can observe a constant blue line which refers to the r2 scores of different models being the same.

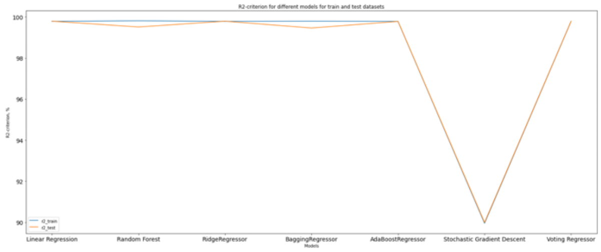


Figure 7: NY taxi data - r2 graph for train and test for different models

3. NYC payroll data: Random Forest and Bagging Regressor performed well among all the other algorithms implemented on this dataset with an r2 score of 99.71 and 99.60 respectively.

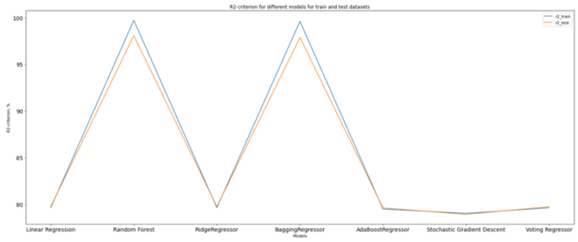


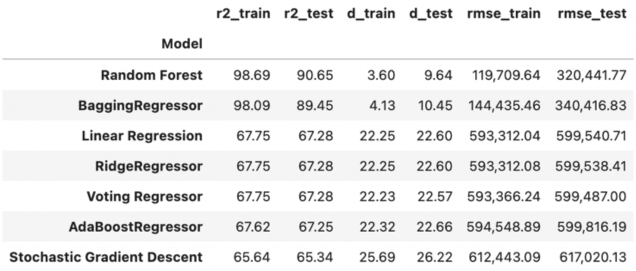
Figure 8: NY payroll data - r2 graph for train and test for different models

G. Discovered Knowledge (Visualization and Interpretation)

From the above analysis, we know what features are important in predicting the values for different use cases and which algorithm performs better to analyze and predict future values. The correlation between different features helps us what feature affects the response variable the most and what features are least effective.

IV. Evaluation

Among all the algorithms implemented on the datasets, Random Forest and Linear Regression performed better than the rest of the algorithms with better accuracy and Stochastic Gradient Descent was the poor performer among all the algorithms.



Model	r2_train	r2_test	d_train	d_test	rmse_train	rmse_test
Random Forest	98.69	90.65	3.60	9.64	119,709.64	320,441.77
BaggingRegressor	98.09	89.45	4.13	10.45	144,435.46	340,416.83
Linear Regression	67.75	67.28	22.25	22.60	593,312.04	599,540.71
RidgeRegressor	67.75	67.28	22.25	22.60	593,312.08	599,538.41
Voting Regressor	67.75	67.28	22.23	22.57	593,366.24	599,487.00
AdaBoostRegressor	67.62	67.25	22.32	22.66	594,548.89	599,816.19
Stochastic Gradient Descent	65.64	65.34	25.69	26.22	612,443.09	617,020.13

Figure 9: Used cars dataset results for different ML algorithms

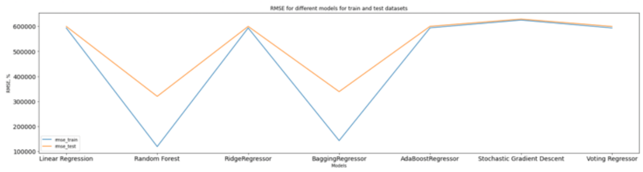
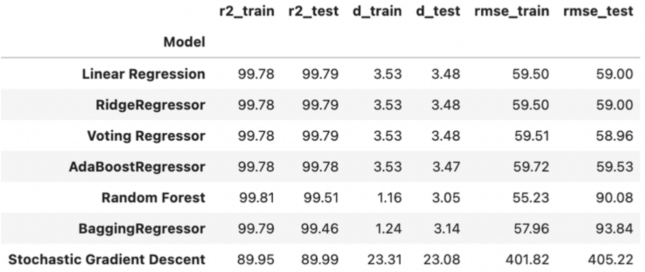


Figure 10: Used cars dataset - RMSE for different ML algorithms

From figure 9, the r2 scores for Random Forest and Bagging Regressor are the best with 98.69 and 98.09 respectively and Stochastic Gradient Descent has the lowest score with 65.64 among all other r2 scores. The RMSE (Root Mean Square Error) for Random Forest is the lowest. We are able to achieve better accuracy for used cars using Random Forest with a low RMSE score.



Model	r2_train	r2_test	d_train	d_test	rmse_train	rmse_test
Linear Regression	99.78	99.79	3.53	3.48	59.50	59.00
RidgeRegressor	99.78	99.79	3.53	3.48	59.50	59.00
Voting Regressor	99.78	99.79	3.53	3.48	59.51	58.96
AdaBoostRegressor	99.78	99.78	3.53	3.47	59.72	59.53
Random Forest	99.81	99.51	1.16	3.05	55.23	90.08
BaggingRegressor	99.79	99.46	1.24	3.14	57.96	93.84
Stochastic Gradient Descent	89.95	89.99	23.31	23.08	401.82	405.22

Figure 11: NY Taxi trip dataset results for different ML algorithms

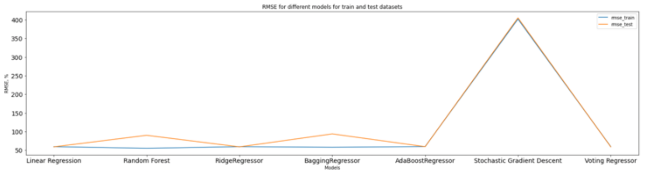
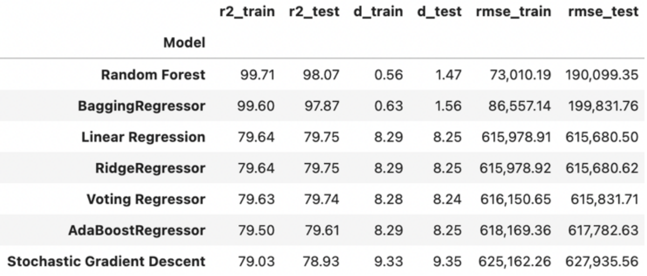


Figure 12: NY Taxi trip dataset - RMSE for different ML algorithms

From figure 10, the r2 scores for Linear Regression, Ridge Regressor, Voting Regressor, and Adaboost Regressor are the best with 99.78 and Stochastic Gradient Descent has the lowest score with 89.95 among all other r2 scores. The RMSE (Root Mean Square Error) for Linear Regression and Ridge Regressor is the lowest. We are able to achieve better accuracy for used cars using Linear Regression with a low RMSE score.



Model	r2_train	r2_test	d_train	d_test	rmse_train	rmse_test
Random Forest	99.71	98.07	0.56	1.47	73,010.19	190,099.35
BaggingRegressor	99.60	97.87	0.63	1.56	86,557.14	199,831.76
Linear Regression	79.64	79.75	8.29	8.25	615,978.91	615,680.50
RidgeRegressor	79.64	79.75	8.29	8.25	615,978.92	615,680.62
Voting Regressor	79.63	79.74	8.28	8.24	616,150.65	615,831.71
AdaBoostRegressor	79.50	79.61	8.29	8.25	618,169.36	617,782.63
Stochastic Gradient Descent	79.03	78.93	9.33	9.35	625,162.26	627,935.56

Figure 13: NY Payroll dataset results for different ML algorithms

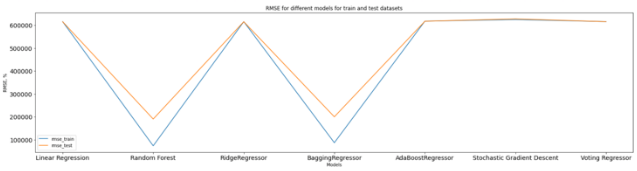


Figure 14: NY Taxi Payroll dataset - RMSE for different ML algorithms

From figure 11, the r2 scores for Random Forest and Bagging Regressor are the best with 99.71 and 99.60 respectively and Stochastic Gradient Descent has the lowest score with 79.03 among all other r2 scores. The RMSE (Root Mean Square Error) for Random Forest is the lowest. We are able to achieve better accuracy for used cars using Random Forest with a low RMSE score.

V. Conclusion And Future Work

From all the implemented algorithms, we can infer that Random Forest and Linear Regression performed better than all other algorithms. However, more algorithms can be implemented on these datasets such as SVM (Support Vector Machines), MLP Regressor, gradient boost with hyperparameter optimization, etc. But due to computational and resource limitations to execute the machine learning algorithms which need high computational power to run. Furthermore, this study can be extended to weather analysis which can help us better understand the pricing of taxi fares and employee behavior at different times of the weather. To support this, we need more data on different domains such as behavior analysis of employees, taxi demand trends, and seasonal trends in taxi usage.

VI. References

[1] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." *Int. J. Inf. Comput. Technol* 4.7 (2014): 753-764.

[2] Pal, Nabarun, et al. "How much is my car worth? A methodology for predicting used cars' prices using random forest." *Future of Information and Communication Conference*. Springer, Cham, 2018.

[3] Wu, Jian-Da, Chuang-Chin Hsu, and Hui-Chu Chen. "An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference." *Expert Systems with Applications* 36.4 (2009): 7809-7817.

[4] Van Thai, Doan, et al. "Prediction car prices using quantify qualitative data and knowledge-based system." *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2019.

[5] Chelliah, Balika J. "Taxi Fare Prediction System Using Key Feature Extraction in Artificial Intelligence." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.6 (2021): 3803-3808.

[6] Chelliah, Balika J. "Taxi Fare Prediction System Using Key Feature Extraction in Artificial Intelligence." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.6 (2021): 3803-3808.

[7] Liu, Ce, and Qiang Qu. "Trip fare estimation study from taxi routing behaviors and localizing traces." *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015.

[8] Bai, Ying-Wen, and En-Wen Wang. "Design of taxi routing and fare estimation program with re-prediction methods for a smart phone." *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*. IEEE, 2012.

[9] Antipova, Anzhelika. "Analysis of Commuting Distances of Low-Income Workers in Memphis Metropolitan Area, TN." *Sustainability* 12.3 (2020): 1209.

[10] Lazar, Alina. "Income prediction via support vector machine." *ICMLA*. 2004.

[11] Bhatia, Komal Kumar. "Prediction Model for Under-Graduating Student's Salary Using Data Mining Techniques." *International Journal of Scientific Research in Network Security and Communication* 6.2 (2018): 50-53.

[12] Kim, Kyeongsu. "Can carsharing meet the mobility needs for the low-income neighborhoods? Lessons from carsharing usage patterns in New York City." *Transportation Research Part A: Policy and Practice* 77 (2015): 249-260.