House Sales in King County, New York City Airbnb, Car Price Prediction using Machine Leaning Models

Etinosa Eghaghe
MSc in Data Analytics
x23138548@student.ncirl.ie

Abstract—This report presents a comprehensive analysis on a collection of three different datasets, utilizing different machine learning Algorithms. The primary goals of the project were to critically compare the performance of different machine learning methods, identify data patterns and reveal insights, knowing the best model for the datasets. Several machine leaning methods were used, including the Knowledge Discovery Database (KDD) approach to achieve good accuracy. The datasets are regarding House sale prediction, Airbnb sales and car price prediction. The approaches involve selecting the dataset, loading, perform pre-processing and cleaning, exploring, transformation, modeling, and evaluation. further sophisticated machine learning analysis can still be done to enhance the method. The R programming language and the robust methodology of multiple linear regression, Lasso regression, K-Nearest Neighbors Algorithm, Radom Forest decision tree modeling was performed in this project. We aspire to extract meaningful insights that not only explain present market trends but also offer a predictive framework for future demand.

Keywords—Machine Learning, Multiple Learning Regression, KDD, Lasso, KNN, Random Forest Regression,

1. INTRODUCTION

1.1 Background

This project explores three datasets, each representing a distinct domain: real estate, hospitality, and sales. The aim of this project is to analyse and model three different datasets: housing sales in King County data, New York City Airbnb data, car sales data were used. This began with loading of the dataset, cleaning and pre-processing the datasets, exploring data distributions, detecting outliers, visualizing correlations, and building predictive models using various techniques such as Multiple Linear Regression, Decision Trees, Random Forests, K-Nearest Neighbors Algorithm, and Lasso Regression.

Dataset Sources:
Dataset1:https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/code
Dataset2:https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data/code
Dataset3:https://www.kaggle.com/datasets/CooperUnion/cardataset/code

1.2 Motivation for the Project

In today's world, understanding the factors that impact housing prices in the real estate sector is paramount. Embarking on the buying or selling a house is an exciting and transforming experience. This machine Learning analysis focuses on understanding the factor influencing housing sale prices, identifying data patterns and insights. The value for house depends on so many factors, the location, year built, how good the condition is overall, grade ,total level in the house, square footage of the home, latitude and longitude coordinate, living room area, number of bedrooms, number of bathrooms and other attributes.it is difficult know all information when buying or selling a house, so a machine learning model to predict the price can help to decide where to buy or sell a house base on the result from the model.

Due to high population and migration Airbnb has become more relevant in our society, with over 1.5 billion guest arrivals in almost every country around the world. Housing challenges in many big cities in recent years has make Airbnb a good place to go for both renter and property owners, renting price are influence by different factor which include the location of the property, type of accommodation, availability of the property etc.
The Airbnb has experience high growth over the years.

In the Ireland not everywhere have regular public transport, for example in Lust where I leave, there is no public transport at mid night., for people that work night and getting tax or renter car are expensive and sometime scarce, people most rely on their vehicle, Knowing the price to buy or sale car is generally very difficult. So, understanding some features in this regard is paramount. The value depends on many factors. Which can include make of the car, model, size, engine, and other attributes get all the information is not an easy task for the buy, the seller also needs to set a good and reasonable price to get profit. Due to all this Machine Learning model to predict prices can help to make decision on buying or selling car.

Research Question(s):

(i) How do various factors, influence house, Airbnb, and car prices?

(ii) What insights can be extracted from the analysis to guide decision-making in buying or selling houses, renting properties on Airbnb, and pricing cars?

(iii) How does the performance of different machine learning models vary across diverse domains, including real estate, hospitality, and automotive sales?

(iv) How do machine learning models, such as Multiple linear regression, Decision Tree, Random Forest, Lasso and K-nearest Neighbour perform in predicting prices?

(v) What insights can be extracted from the analysis to guide decision-making in buying or selling houses, renting properties on Airbnb, and pricing cars?

2 RELATED WORKS

2.1 House Sale in King County:

Previous research on house sales has explored various factors influencing property prices [2], including latitude and longitude, size condition and so on.

Notable studies by Dubin [3] and [4] have utilized advanced regression models to predict house prices. The work in [4] introduced a novel approach incorporating geographical features to improve prediction. While this represents a significant advancement, there is a need for further investigation into generalizability of such models across diverse neighborhoods in king county.

Previous studies have often focused on specific neighborhoods, potentially limiting the applicability of their findings to the entire county. A more comprehensive analysis across diverse regions is necessary. The usefulness of prior work lies in establishing foundational insights into the housing market dynamics in King County. However, limitations in feature selection and model interpretability warrant further exploration.

Datasets such as [Dataset1] and [Dataset2] have been widely employed in King County housing studies. Methodologies include linear regression, decision trees, and ensemble methods. While some studies have successfully identified the impact of location and amenities, there is room for more nuanced analyses to better capture the intricacies of the housing market.

2.2 New York City Airbnb:

Previous research on Airbnb pricing and popularity in New York City has addressed the complex task of determining optimal listing prices and factors influencing booking rates [5]. Researchers have leveraged datasets containing listing details, customer reviews, and geographical information.

Studies by [6] and [7] have applied machine learning models to predict Airbnb prices. These studies provide valuable insights, but their methodologies could benefit from a more extensive consideration of temporal factors, such as seasonality and special events.

[6] explored the impact of neighborhood characteristics on Airbnb pricing, offering valuable insights into the spatial dynamics of listing prices. However, limitations in data collection methods may impact the generalizability of findings.

The limitation of some studies in accounting for temporal dynamics in pricing suggests the need for more sophisticated models that consider fluctuations in demand over time.

Despite limitations, previous work has laid the foundation for understanding the role of location, listing attributes, and customer reviews in determining Airbnb prices in NYC.

Datasets like [Dataset3] and [Dataset4] have been instrumental in Airbnb studies. Machine learning methods, including regression and clustering, have been commonly employed. The variability in methodologies highlights the ongoing exploration of effective approaches to predict Airbnb prices in the dynamic NYC market.

2.3 Car Price:

Previous research [8] on predicting car prices using machine learning has focused on understanding the factors influencing the market value of automobiles. Researchers have employed datasets containing details such as brand reputation, model specifications, and historical sales data.

In the book, regression models were employed to predict car prices based on brand reputation and technical specifications. However, the study lacked a thorough exploration of the impact of economic variables such as fuel prices and inflation on car prices. [9] introduced a comprehensive model incorporating economic factors but may benefit from a more in-depth analysis of feature importance and model interpretability.

Previous studies have made significant strides in predicting car prices, but the limited consideration of economic variables suggests an opportunity for more holistic models.

The usefulness of prior work lies in providing foundational insights into the impact of brand reputation and technical specifications on car pricing.

Datasets used in [10] and [11] have been employed in car price prediction studies. Regression models and ensemble methods have been commonly used. The application of machine learning models to predict car prices showcases the potential for leveraging historical data to inform decision-making in the automotive market.

[Author8] introduced a comprehensive model incorporating economic factors but may benefit from a more in-depth analysis of feature importance and model interpretability.

Previous studies have made significant strides in predicting car prices, but the limited consideration of economic variables suggests an opportunity for more holistic models.

The usefulness of prior work lies in providing foundational insights into the impact of brand reputation and technical specifications on car pricing.

Datasets such as [Dataset5] and [Dataset6] have been employed in car price prediction studies. Regression models and ensemble methods have been commonly used. The application of machine learning models to predict car prices showcases the potential for leveraging historical data to inform decision-making in the automotive market.

3 METHODOLOGY

For the three datasets on this project, KDD methodology is deployed for data mining. Similar steps are being taken across the three datasets it includes the following.

i)Data Gathering: At this stage data is imported and a target variable is identifier to be used for execution of KDD.

ii) **Data preparation:** we apply **EDA** to investigate the data to build a strategy for each variable and to get a good understanding of while selecting a which machine learning model.

iii) **Data Transformation:** This involves the process of adapting the data to align with the specific requirement of the model being used, during this phase the data is modified to suit the application of the model.

iv) **Data Mining:** this involves looking for patterns based on our prediction goal.

v) **Result and Evaluation:** The stage involves pattern been used to generate inferences.

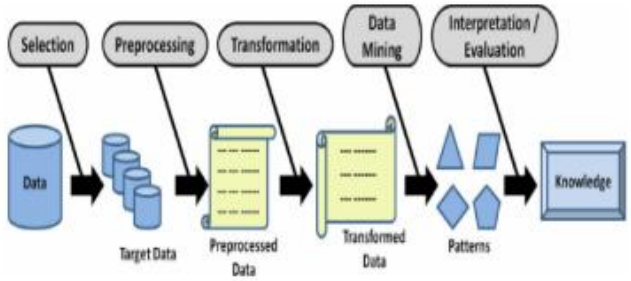

Figure 1: KDD Methodology

4 DATA EXPLORATION, CLEANING AND TRANSFRMATION

4.1 Dataset 1: Housing sales

4.1.1 Data Overview and Exploration

The dataset originally had 21613 rows and 21 columns, which implies that the dataset consists of 21 information of house features in King County. The features are:

Id, date, price, Bedrooms, sqft living, sqft lot, Floors, Waterfront, View, condiction, Grade, sqft above, sqft basement, yr built, yr renovated, Zipcode, lat, log, sqft living15, sqft lot15

Dataset Loading: The initial steps involve sting the working directory and loading the dataset. Examine the structure of the dataset through the following practices.

- **Summary()** : this function show summary for each variable, such as mean, median, mode, quarties.
- **Dim()** : used to view the dimenion of the dataset
- **Sapply()**: This function type of variables in the dataset, which contians numerical and categoriacl variables.
- **Str()** : This function displays an overview of the dataset structure.

4.1.2 DATA PREPROCESSING

4.1.2a Handling Missing Values: This stage missing values was check because in data processing handling missing values is import in other to ensure the accuracy and reliability of subsequent analysis [3] Data can be missing due to various reason which include error during data collection or simply the absence of information for certain observations a count of missing values in column was display. We used 'is.na()' function to identify missing values but with the dataset used there were no missing values .Infinite number was also examined.

```
            id           date          price       bedrooms      bathrooms    sqft_living
             0              0              0              0              0              0
      sqft_lot         floors     waterfront           view      condition          grade
             0              0              0              0              0              0
    sqft_above sqft_basement       yr_built   yr_renovated        zipcode            lat
             0              0              0              0              0              0
          long  sqft_living15     sqft_lot15
             0              0              0
```

4.1.2b Outlier Detection and Correction: This is a crucial phase to ensure that models are duly influenced by extreme values. Outliers are data points that significantly differ from most of the data and can have a substantial impact on the results of analysis.[1] Identifying and addressing outliers help improve the robustness and reliability of the models. Boxplots were utilized to visualize and correct outliers using the interquartile range (IQR) method. Before removing the outlier, the dataset had 21613 rows and after remover became 13868 rows. As show in Figure2 an 3 respectively.

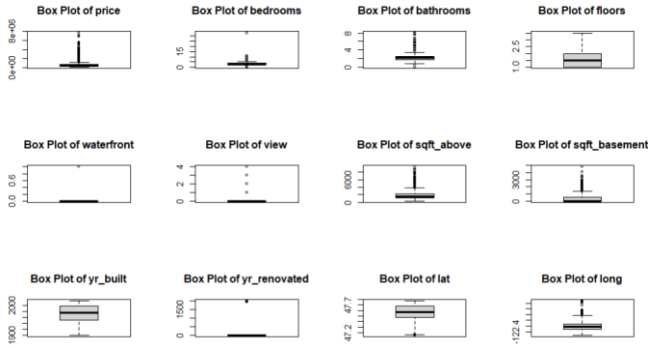

Figure 2: Subset Before IQR Test

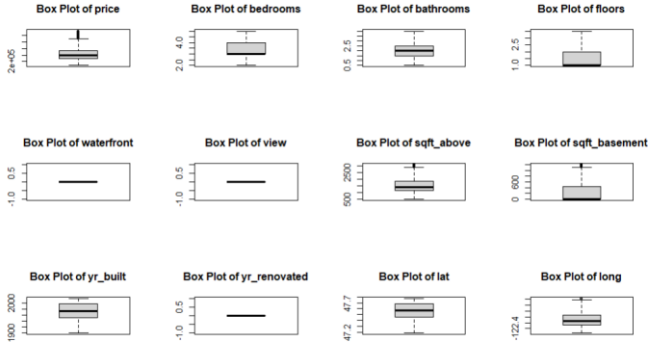

Figure 3: Subset After IQR Test

4.1.3 Feature Selection: This involves choosing a subsect of features the are relevant to retain the most informative and influential features. This leads to eliminating redundant or less important ones, it is difficult due to many reasons. But it improves model performance, reduce overfitting, computational efficiency and enhance interpretability, for this analysis unnecessary columns that has no range and columns that are not correlated were removed from the dataset to streamline the modeling process. Also to examine the relationship correlation matrix(Figure 4&5) was plotted using heatmap, we find out that variables like waterforont, view,yr_renovated were not correlated so this columns were drop.

waterfront	view	yr_renovated
Min. :0	Min. :0	Min. :0
1st Qu.:0	1st Qu.:0	1st Qu.:0
Median :0	Median :0	Median :0
Mean :0	Mean :0	Mean :0
3rd Qu.:0	3rd Qu.:0	3rd Qu.:0
Max. :0	Max. :0	Max. :0

sample of variables with no range

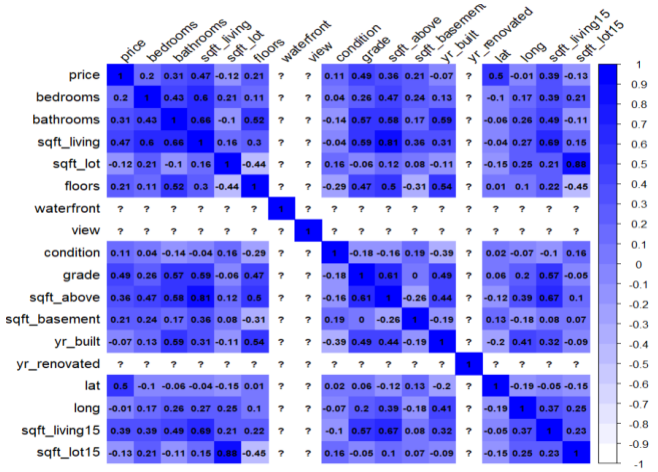

Figure 4: **Correlation Analysis: Heatmap Visualization (Dataset1)**

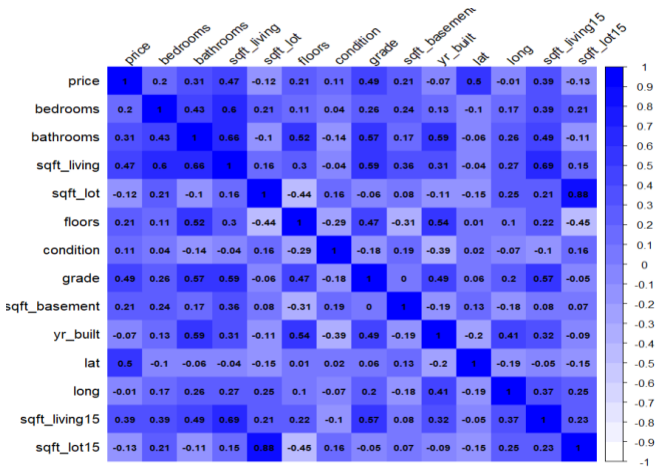

Figure 5: **Correlation Analysis: Heatmap Visualization (Dataset1)**

Standardization: Also known as normalization used to scale numeric variables in a dataset for them to have similar scale, this make sure that variable is on common scale to prevent certain variables from disproportionately influencing the model training process. It transforms each numerical variable in the dataset so that it has a mean of 0 and standard deviation of 1. In this analysis selected numerical features were standardized using min-max normalization to ensure consistent scales.

$$Z = (X - mean(X)) / std(X)$$

Model Selection

For this dataset the aim was to predict the price of house so, price variable was selected as depended variable and the others were used as independent variables.
Three machine learning models were built and hyperparameters were tweaked where applicable. Multiple models were investigated because they have different advantages and disadvantages. The models selected were all the regression type because the desired output 'price' is continuous number. Below are the three models.

1. Multiple Linear Regression
2. Decision Tree
3. Random Forest

4.2 Dataset 2: NYC Airbnb

4.2.1 Data Overview and Exploration

The NYC Airbnb data contains 48895 rows and 16 columns originally, the price variable as the dependent variable and other variable like id, host_id, neighbourhood_group, neighbourhood, latitude, longitude, room-type and so on, as independent variables.
This data set was processed similarly to the first data set, which involves. Loading the data, checking the summary, the shape of the data and displaying an overview of the data.

4.2.2 DATA PREPROCESSING

4.2.2a Handling Missing Values: A count of missing values in column was displayed. Using the appropriate function ('is.na()') to identify missing values. In column reviews_per-month 10052 missing, name 16 missing , host_name 21,and last_view 10052 missing values but for this analysis missing values were handled by inputting with the mean of the non-missing values in reviews_per-month column and colunms name, host_name, last_view were dropped because not needed.

room_type	price
0	0
minimum_nights	number_of_reviews
0	0
reviews_per_month	calculated_host_listings_count
10052	0

Subset of missing values viewed.

4.2.2b Data Encoding

Some of the columns which include neighbourhood_group, neighbourhood,and room_type were not numerical variables, so were covered to numerical type by using label encoding and the original factor variable were removed from the dataset.

4.2.2c Outlier Detection and Correction: Boxplots were utilized to visualize and correct outliers using the interquartile range (IQR) method to improve the robustness and reliability of the model. Before removing the outlier, the dataset had 48895 rows and after remover became 23262 rows. Seen in Figure 6 and 7 below.

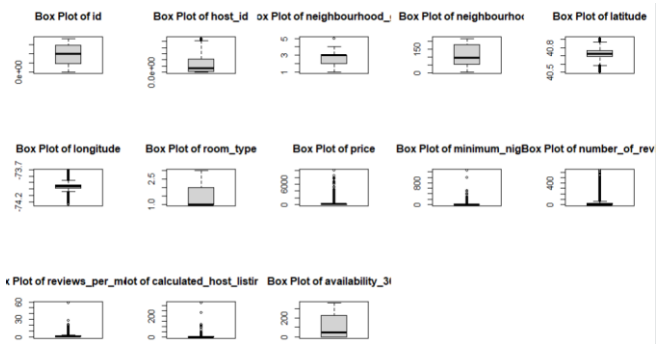

Figure 6: Subset Before IQR Test (NYC Airbnb)

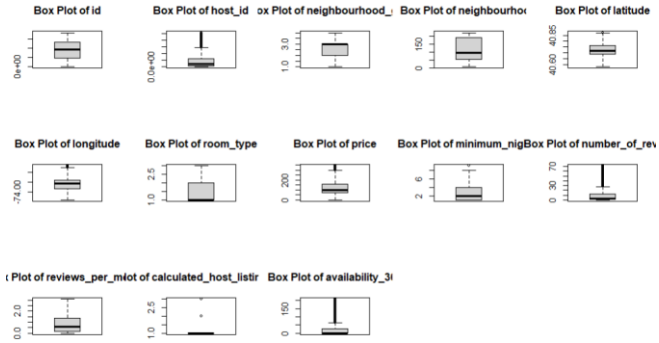

Figure 7: Subset After IQR Test (NYC Airbnb)

4.2.3 Feature Selection: correlation matrix was plotted to examine the relationship between variables to avoid multicollinearity in the model.

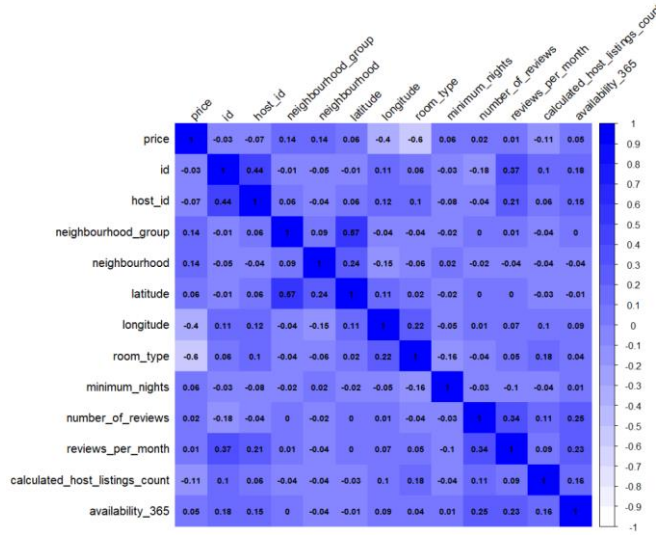

Figure 7: Correlation Analysis: Heatmap Visualization (Dataset2)

Standardization: similarly to the first dataset selected numerical features were standardized using min-max normalization to ensure consistent scales.

$$Z = (X - mean(X)) / std(X)$$

Model Selection: The NYC Airbnb price was also selected as the dependent variable and accordingly other variable in the dataset were used as independent variables, Lasso regression applied to build model and evaluated.

4.3 Dataset 3: CAR PRICE PREDICTION

4.3.1 Data Overview and Exploration

The 11914 rows and 16 columns originally, the 'price' variable as the dependent variable and other variables as independent variables listed as follow:

Make, Model, Year, Engine_FuelType, Engine_HP ,Engine_Cylinders ,Transmission_Type ,Driven_Wheels ,Number_of_Doors, Market_Category, Vehicle_Size, Vehicle_Style, highway_MPG, city_mpg, Popularity.

The process began with Loading the data, checking the summary, the shape of the data, displaying an overview of the data., renaming of some columns, checking, and handling missing values, examine infinite number, visualization using box plots for numerical variables and bar plot for categorical variables, factor variables were converted to numerical variables. Outliers were handled.

4.3.2 DATA PREPROCESSING

4.2.2a Handling Missing Values: This stage missing values was checked and there were missing values in three columns which in Engine_Fueltype,(3missing) Engine_HP(69 missing) and Engine_cylinders(30missing).. we decided to remove the rows with the missing values since it was not much. After reomoving the missing values, the dimension of the dataset was reduced to 11812 rows. Also, infinite numbers were examined. Ggplot2 was used to plot the percentage of missing value as shown below.

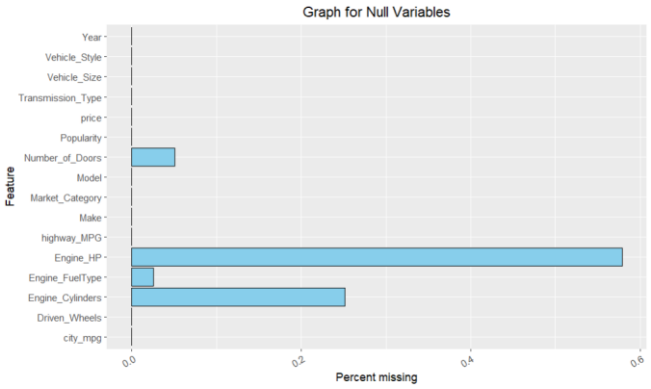

Figure 9: Null Variable % plot

Visualization:

Histogram: this allows us to visualize each variable for a better understanding of the data distribution and potential patterns. Geom-histogram is used for numerical variables and geom-bar for categorical variable, distribution view.

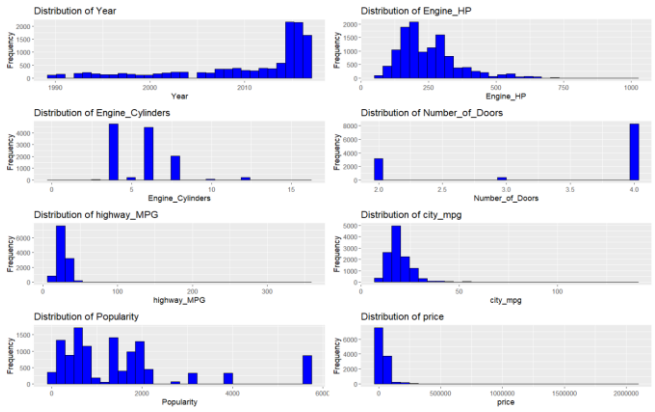

Figure 10: Plot Numeric Variables

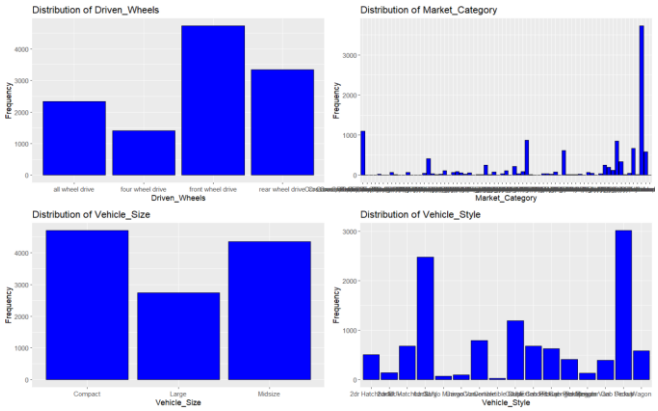

Figure 11: Plot Subset Factor Variables

4.2.2b Data Encoding

Label encoding method is applied to convert Factor variables (make,model,Engine_FuelType, etc) at this stage to numeric variable, making them suitable for the machine learning algorithm and the original factor column removed because no longer needed.

4.2.2c Outlier Detection and Correction: Boxplots were also utilized to visualize and correct outliers using the interquartile range (IQR) method. Before removing the outlier, the dataset had 11812 rows and after remover became 8718 rows. The correlation matrix to visualize the relationship between variables was plotted.

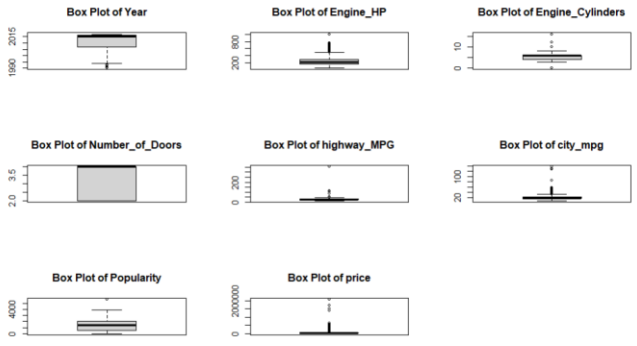

Figure 12: Subset Before IQR Test (Car Prediction)

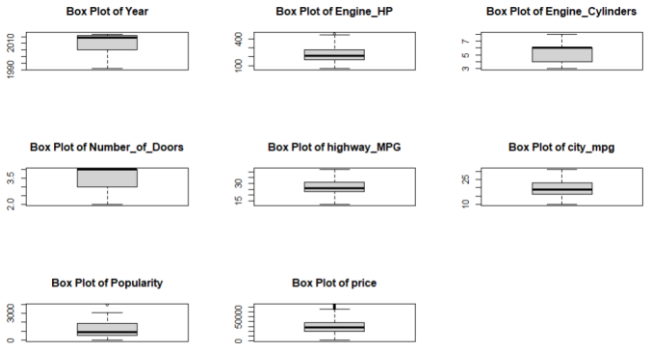

Figure 13: Subset After IQR Test (Car Prediction)

4.2.3 Feature Selection: correlation matrix was also plotted to examine the relationship between variables to avoid multicollinearity in the model.

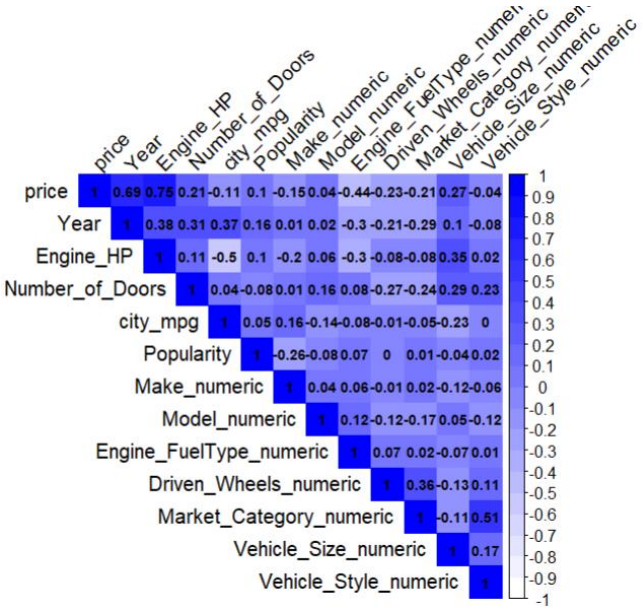

Figure 14: Correlation Analysis: Heatmap Visualization (Dataset3)

Model Selection: In the Car dataset, price was also selected as the dependent variable and accordingly other variable in the dataset were used as independent variablesA correlation matrix was also plotted to observe the collinearity between the independent variables. For further analysis, this was crucial to identify potential violations when dealing with model evaluation. KNN regression a model developed and evaluated.

5 MODEL AND EVALUATION

5.1 Dataset 1: Housing sales prediction

Model Development:

The clean data at this stage is split into training and testing dataset, by setting the seed to generate randomly, to fit the models and predictions are made. This separation helps to ensure that the model is tested on data it has never seen during training, providing a more realistic assessment of its generalization performance. For this analysis we split on a ratio of 80% for training and 20% for testing which result to a dimension of

Train Data (11469, 14)

Test Data: (2399, 14)

5.1.1 Model 1: Multiple Linear Regression

The MLR model was developed using price variable as the dependent variable and selected independent variables. Model summary is generated to provide a comprehensive overview of the performance. Key statistics below show the residual standard error, R-squared, adjusted R-squared, F-statistic, and P-value. These show the goodness of the fit to training data of the model.

- Model performance was assessed using summary.
- Variance inflation (VIF) for multicollinearity to see the linearity.
- Non-constant variance test, and autocorrelation test (Durbin-Watson).
- Histogram of residuals provided insight into model fit.

Model Statistic	Value
Residual standard error	0.229
Degree of freedom	11455
Multiple R-squared	0.6881
Adjusted R-squared	0.6877
F-statistic	1944
F-statistic degree of freedom	13 and 11455
P-value	< 2.2e-16

Summary (MLR)

Chi-square	Degree of Freedom (DF)	P-value
5.980261	1	0.014467

Non-constant Variance Score Test (MLR)

lag	Autocorrelation	D-W Statistic	p-value
1	0.002281887	1.995262	0.842

DurbinWastonTest (MLR)

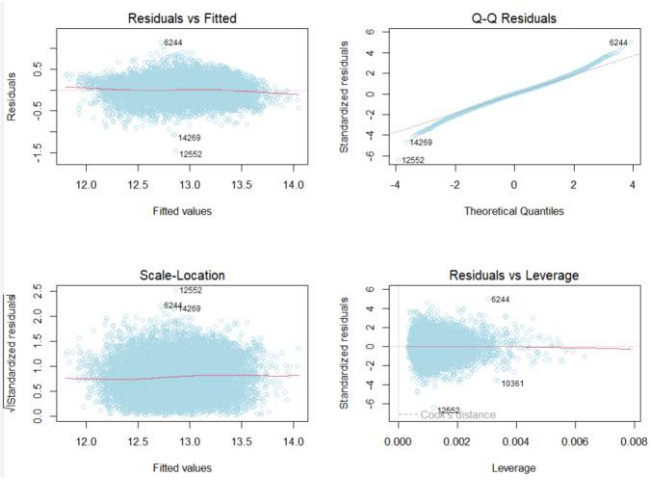

Figure 15: relationship b/w variable

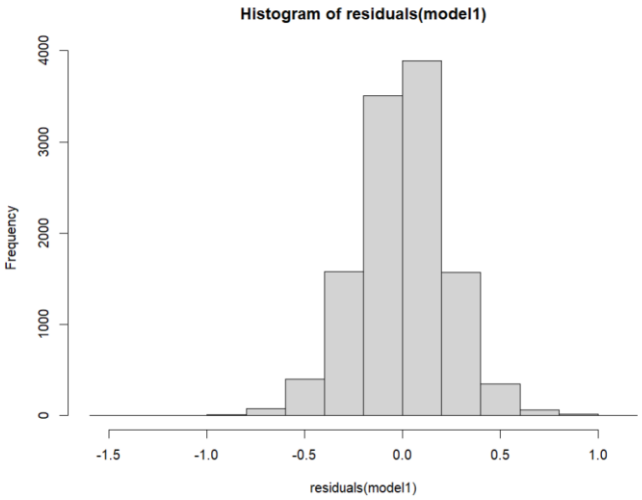

Figure 16: Histogram of Residuals

Cross Validation

At this stage the multiple linear regression model was cross-validated using 5-fold cross-validation and the result of the performance metrics indicate how well is the model expected to perform on unseen data, this report provides insight into the accuracy and reliability of the model prediction. Below are the results:

RMSE: 0.2290382
Rsquared: 0.6875987
MAE: 0.1770411

5.1.2 Model 2: Decision Tree Regression

This was developed and tested. Visually inspected. The R-squared score: 62.63834 and a cross validated using 5-fold cross-validation also.

cp	RMSE	Rsquared	MAE
0.06135571	0.2730705	0.5553324	0.2128188
0.09164787	0.2993881	0.4653661	0.2359638
0.43271539	0.3497692	0.4280267	0.2811092

Cross-validation (Decision Tree)

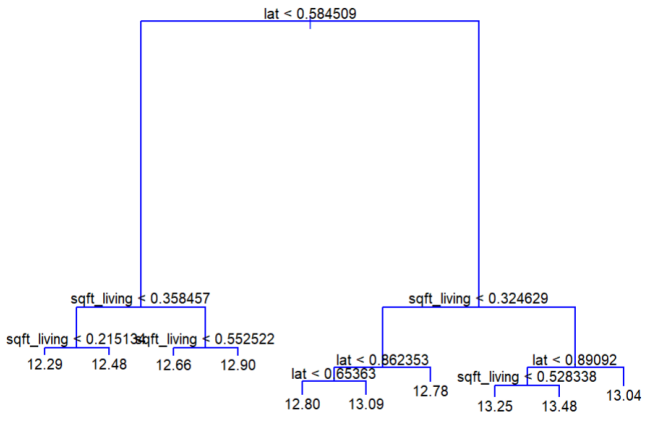

Figure 17: ROC Curve plot (Decision Tree)

5.1.3 Model 3: Random Forest Regression

The Random Forest Regression model was developed and validated. Visually inspected. The R-squared score: 79.78922 and a cross validated

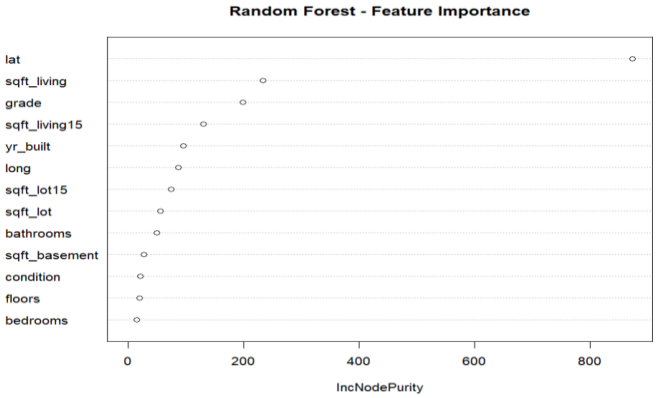

Figure 18: Random Forest Plot (Feature Importance)

```
Random Forest

11469 samples
   13 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 9176, 9175, 9175, 9175, 9175
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE
   2    0.1688481  0.8372658  0.1221646
   7    0.1565203  0.8546007  0.1121062
  13    0.1574910  0.8524988  0.1131496

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 7.
```

Random Forest cross validation Result

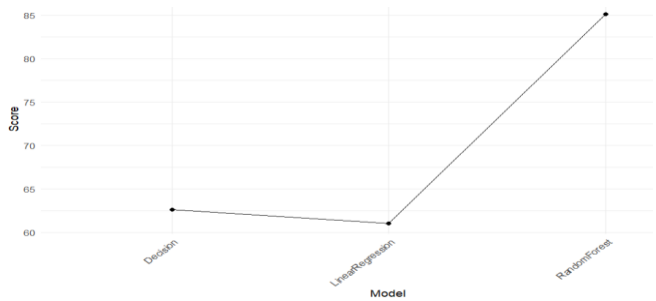

Figure 19: R-Squared Result (MLR, DTR,RF)

5.2 Dataset 2: NYC Airbnb

Like the first dataset, the data split into training and testing dataset, by setting the seed to generate randomly, to fit the models and predictions are made. For this analysis we split on a ratio of 80% for training and 20% for testing which result to a dimension of

Train Data (18616, 13)

Test Data: (4646, 13)

5.2.1 MODEL 1: Lasso Regression

Lasso regression model was created using glmet library. Apha=1 and lamda = 0.001. The model was train, prediction on test data, calculated residual using response variable, plot of residuals and the mean squared error, R2 score and mean absolute error obtain.

MSE: 0.140624

R2: 46.57698

MAE: 0.1047756

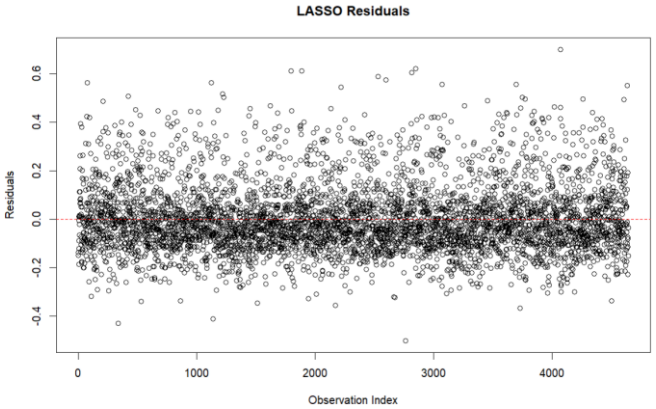

Figure 20: Lasso Regression Residual

5.2.2 MODEL 2: Lasso Regression

The second lasso model was developed. A model matrix for both training and testing data was created, cross validation set with 10 number of folds The cross-validation result obtains. Shown below.

```
Measure: Mean-Squared Error

      Lambda Index Measure      SE Nonzero
min 0.000273    66 0.02013 0.0002493      10
1se 0.005353    34 0.02038 0.0002284       8
```

Best lambda value obtains from cross validation (0.0002726881) and lasso model was fit using the best lambda. The residual was calculated and plot.

MSE: 0.140592, R2: 46.59862, MAE: 0.1047279

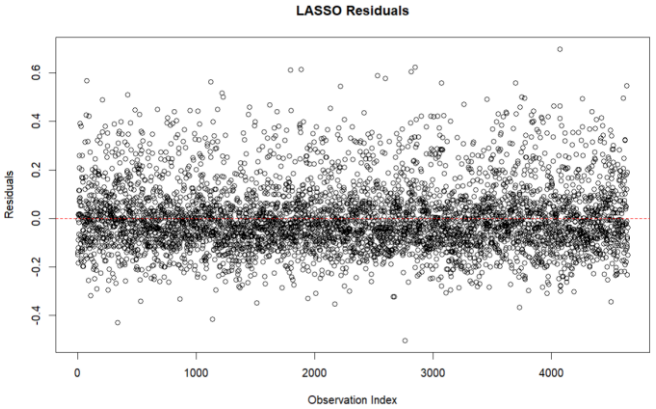

Figure 21: Lasso Regression Residual

5.3 Dataset 3: Car Price Prediction

Also, the data split into training and testing dataset, by setting the seed to generate randomly, to fit the models and predictions are made. For this analysis we split on a ratio of 80% for training and 20% for testing which result to a dimension of

Train Data (6974, 13)

Test Data: (1744, 13)

5.3.1 MODEL: KNN Regression Model

Knn regression model was created using class library. Beginning with model training and Hyperparameter turning, plotting RMSE for different k values, making predictions

with the best k, model evaluation, plotting prediction vs. actual prices, plotting accuracy vs. Neighbors, and finally printing overall accuracy.

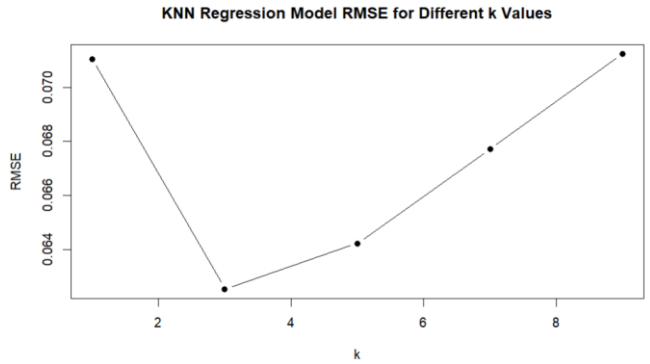

Figure 22: KNN Regression Model RMSE for Different k values

This line plots is to visualize the how the performance changes with k. we can see that the best is around k-5
MSE: 0.01016505
R2: 79.86067
MAE: 0.05586742

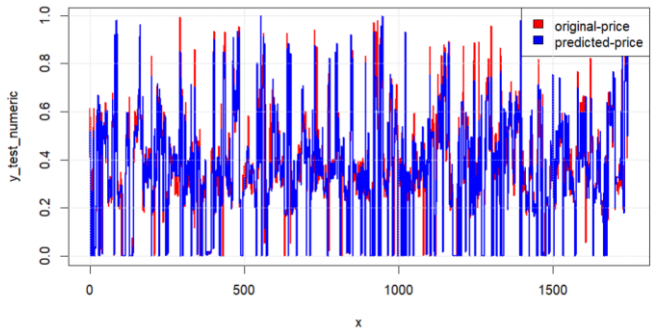

Figure 23: Car Test Data Prediction

MLRegresion	**61.05415**
DecisionTree	**62.63834**
Random Forest	**79.78202**

Model R-Squared

6 CONCLUSION AND FUTURE WORK

In this comprehensive analysis, three distinct datasets were explored and analyzed using various machine learning algorithms to predict house sale price, Airbnb price, and car prices. Multiple linear regression, decision tree regression, random forest regression, lasso regression, and K-nearest neighbor regression. were employed to build models for each dataset, and models were evaluated using relevant performance metrics.

For the housing sale dataset, the random forest regression model demonstrates the highest R-square value (79.79%), indicating its higher predictive performance compared to multiple linear regression with (R2:61.05%) and decision tree regression(R2:62.64%) models. The analysis focused on understanding the factors influencing house prices in King County, with emphasis on location, size, condition, and other attributes.

In the case of the New York City Airbnb dataset, lasso regression models were employed, and the best lambda value was determined through cross-validation. The models achieved an R-squared value of approximately 46.6%, providing insights into the factors affecting Airbnb rental prices, such as neighborhood, room type, and other relevant features.

The car price prediction dataset utilized K-nearest neighbors (KNN) regression, and the model achieved an R-squared value of 79.86%. The analysis considered factors like make, model, year, engine specifications, and more to predict car prices accurately.

While the current analysis provides valuable insights into predicting prices for houses, Airbnb, and cars, there are opportunities for further enhancement and exploration which can include:
Incorporating Additional Features: Including more features or external factors could improve the accuracy of predictions. For example, economic indicators, neighborhood characteristics, or seasonal variations might contribute to a more comprehensive understanding.
Fine-tuning Models: Hyperparameter tuning for machine learning models could be further explored to optimize model performance.
In conclusion, this project saves as groundwork for predicting prices in diverse domains using machine learning algorithms. Further techniques and exploration can enhance the performance and applicability of these models for real world decision making.

7 REFERENCE

[1] L. M. Hoffman and B. S. Heisler, Airbnb, Short-Term Rentals, and the Future of Housing. Taylor & Francis, 2020.
[2] J. Albert and J. Hu, Probability and Bayesian Modeling (Chapman & Hall/CRC Texts in Statistical Science). CRC Press, 2019.
[3] R. A. Dubin, Predicting House Prices Using Multiple Listings Data. SSRN, 2001.
[4] J. E. Burt, G. M. Barber, and D. L. Rigby, Elementary Statistics for Geographers. Guilford Publications, 2009.
[5] J. A. Oskam, The Future of Airbnb and the 'Sharing Economy': The Collaborative Consumption of our Cities (The Future of Tourism). Channel View Publications, 2019.
[6] S. Sedkaoui and M. Khelfaoui, Sharing Economy and Big Data Analytics. Wiley, 2020.
[7] S. Rey, D. Arribas-Bel, and L. J. Wolf, Geographic Data Science with Python (Chapman & Hall/CRC Texts in Statistical Science). CRC Press, 2023.
[8] N. Chaki, N. D. Roy, P. Debnath, and K. Saeed, Proceedings of International Conference on Data Analytics and Insights, ICDAI 2023 (Lecture Notes in Networks and Systems). Springer Nature Singapore, 2023.

[9] M. M. Satapathy and M. R. Prust, PROCEEDINGS OF ACADEMICS WORLD INTERNATIONAL CONFERENCE. Institute for Technology and Research, 2022.
[10] J. Maindonald and J. Braun, Data Analysis and Graphics Using R: An Example-based Approach (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, 2006.
[11] M. B. Ahmed, A. A. Boudhir, D. Santos, R. Dionisio, and N. Benaya, Innovations in Smart Cities Applications Volume 6: The Proceedings of the 7th International Conference on Smart City Applications (Lecture Notes in Networks and Systems). Springer International Publishing, 2023.
[12] S. García, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining (Intelligent Systems Reference Library). Springer International Publishing, 2014.