Predicting Prices of Used Cars

Introduction

This project is about predicting the price of used-cars based on 12 independent variables such as engine capacity, mileage, age, location and so on. The dataset was provided from kaggle website and the data was collected from cars in different cities of India. This project was mainly performed for training and learning purposes. The results of this project can be used to define the most important parameters on depreciation of used-car prices.

Data Cleaning

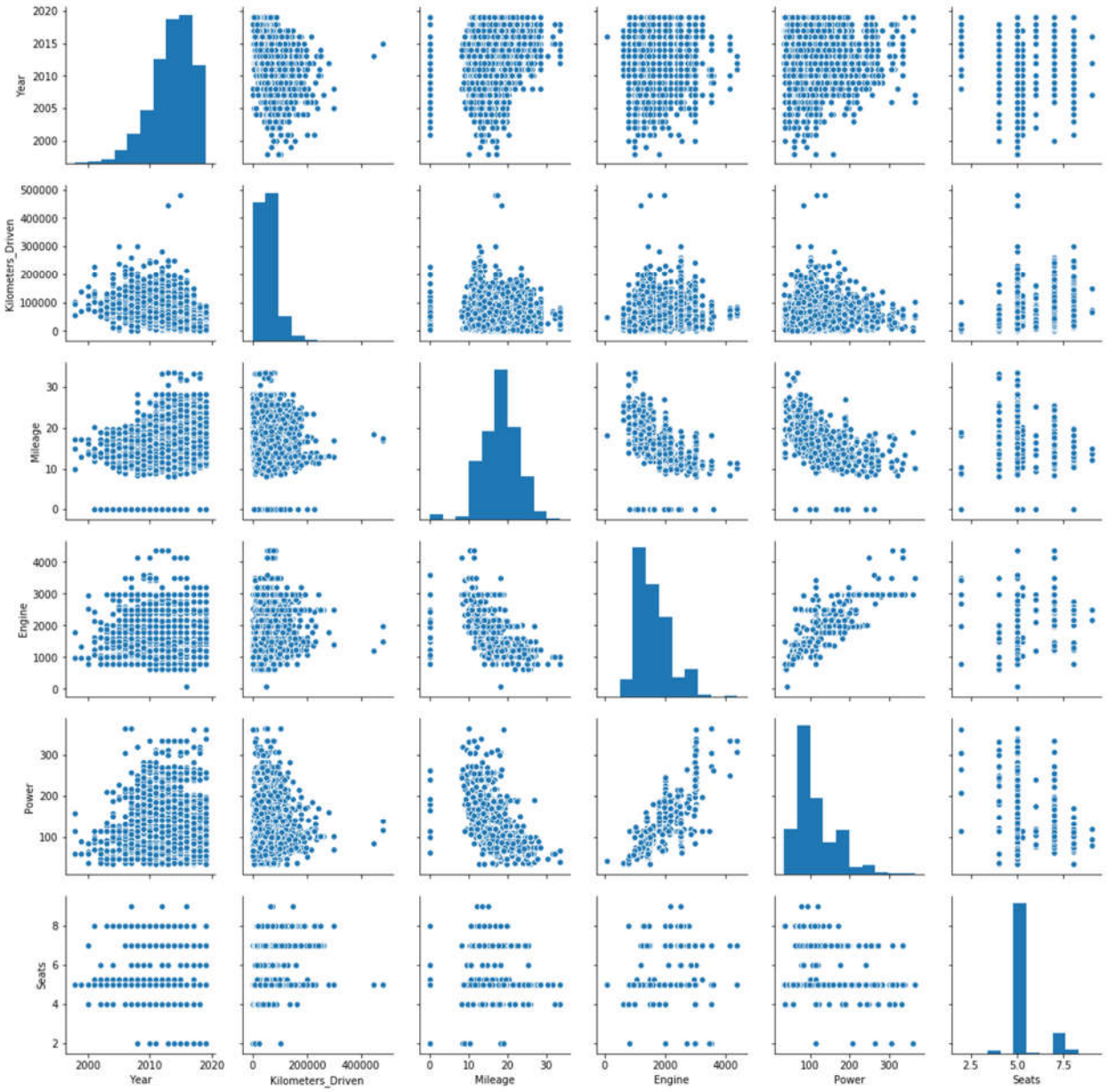
The data cleaning part of the project contained four steps: fixing corrupt data, replacing missing values, removing the outliers, and fixing wrong data types.

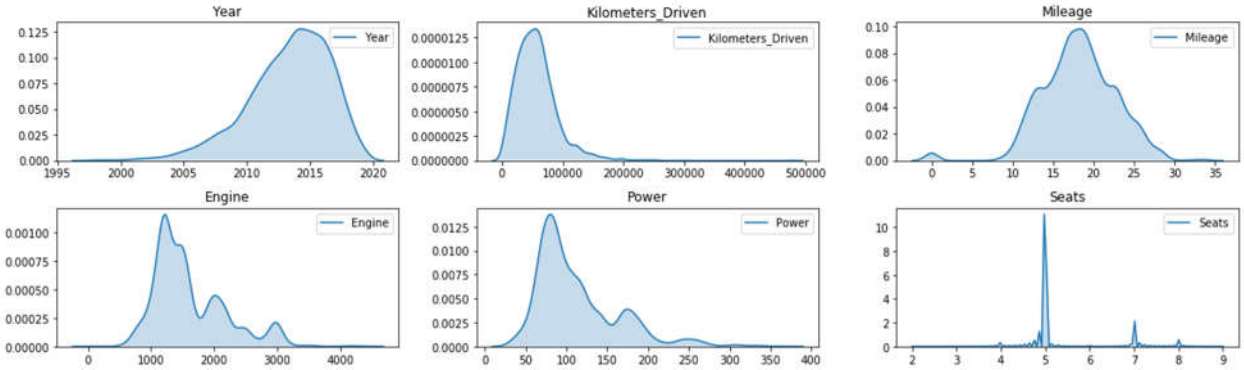
A brief summary of the dataset showed that some features have missing values (i.e., Mileage, Engine, Power, Seats, and New_Price). The 'New_Price' column is missing about 90% of its data, so that I did not use it for data modeling as it could not add much information to the model. Some numerical columns had an 'Object' format (i.e., Mileage, Engine, and Power) as their values were recorded along with their measurement units. I removed the unit part and converted the variables to ‘numeric’ format.

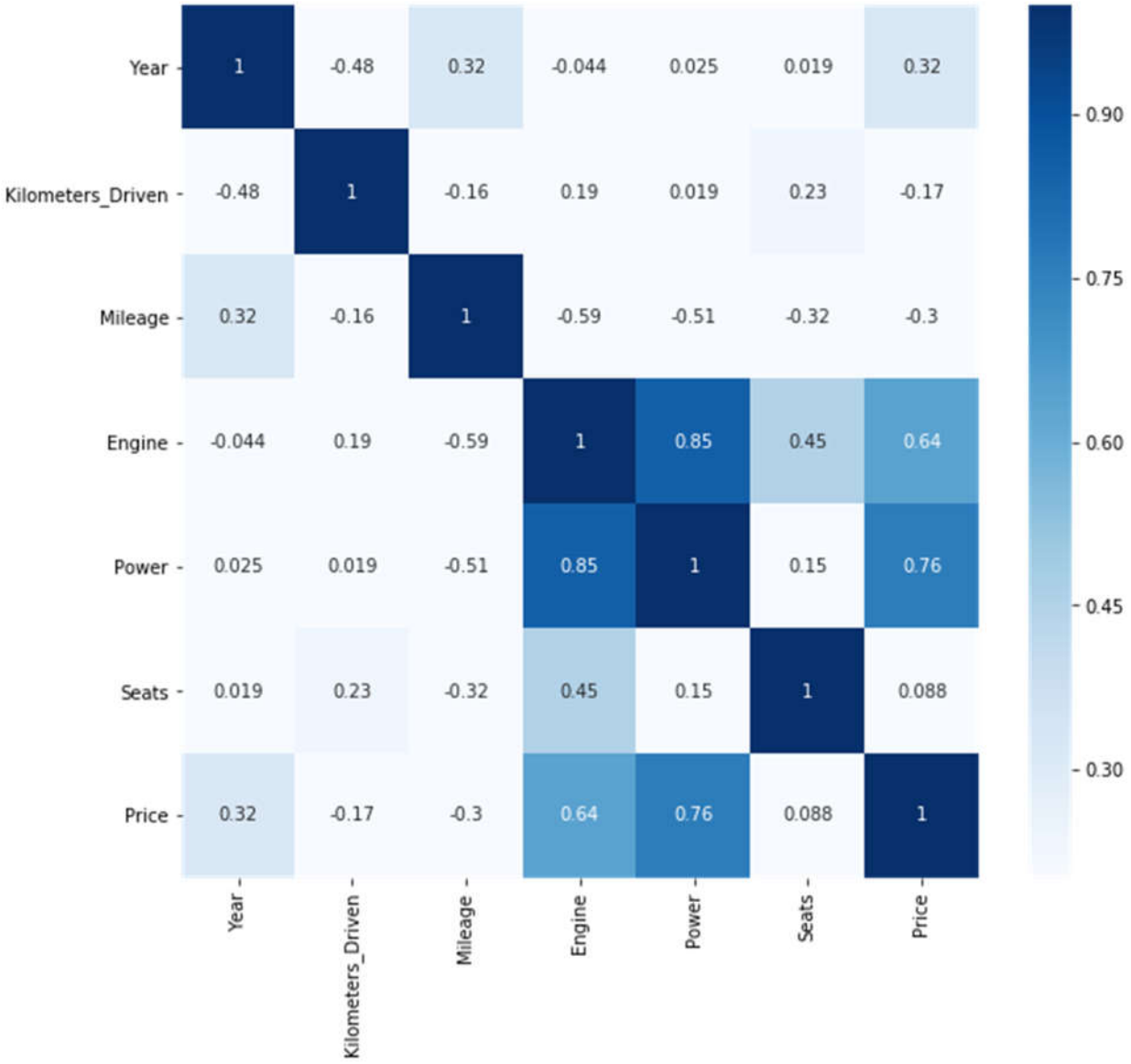
Some features contained ‘null’ or 0 values and I replaced these values with ‘nan’ to make them recognizable and consistent with the remained missing values. To address the missing values, I firstly removed the features which more than 60 percent of their data were missing. For the sake of simplicity, I replaced all the missing values of the numerical features with their mean values, and for the categorical features with their most frequent values. I removed some of the datapoints which contained outlier values. I considered outliers to be out of the range of five standard deviation from the mean value. I used this wide margin to make sure that I don’t loose much data.

Data Exploration

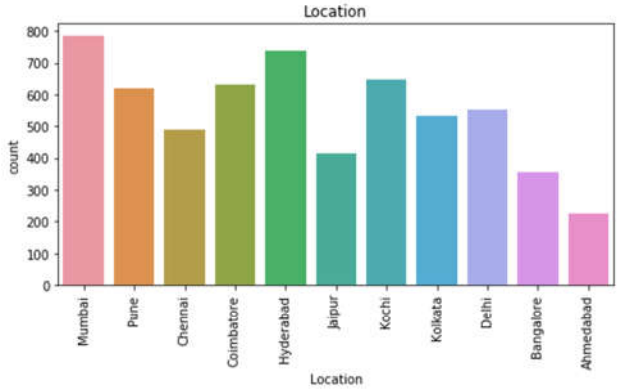
I used descriptive statistics and data visualization for this phase of the project. I utilized pairplots and heatmap for numerical features to demonstrate the correlation between each pair of features or the target variable. I also used kdeplot to show the distribution of numerical variables. The graphs indicated that ‘Mileage’ values were normally distributed, the target variable and the ‘Kilometers_Driven’ variable were highly right skewed, and the ‘Year’ variable was mostly skewed to the left. The ‘Seats’ variable was mostly reported to be 5 and had nearly normal distribution. According to the heatmap, 'Power' and 'Engine' parameters were highly correlated and were skewed to the right. These two parameters also showed highest effects on the 'Price' among all numerical features.

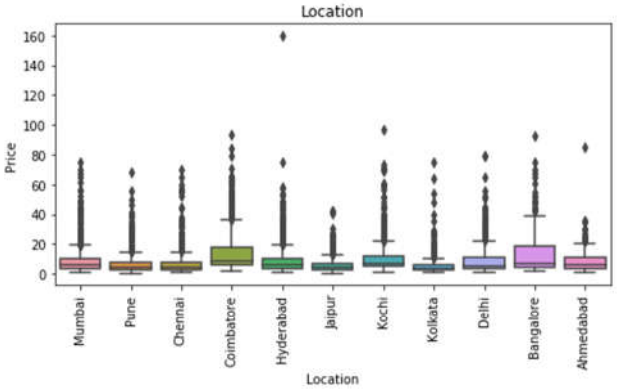

Year
Kilometers_Driven
Mileage
Engine
Power
Seats

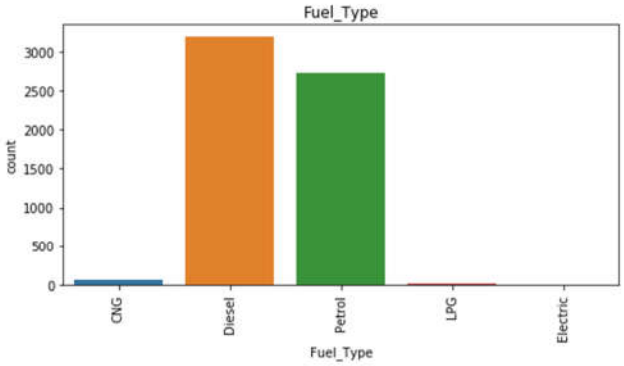

Year
Kilometers_Driven
Mileage
Engine
Power
Seats

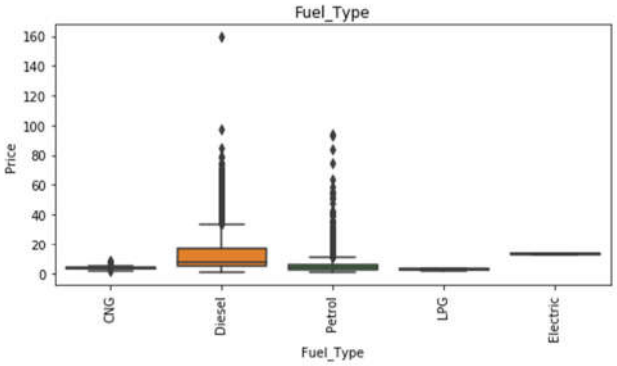


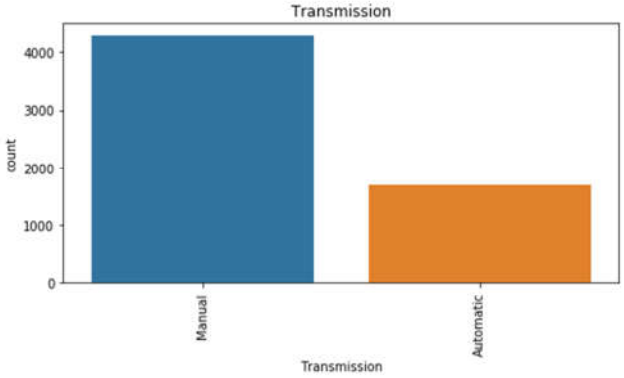
The distribution of categorical variables was demonstrated by barplot and their relation to the target variable was shown using boxplots. Some of the categorical variables had more than 10 levels (unique values) whereas some others had a few numbers of unique values. Moreover, some categorical levels contained a few numbers of datapoints. These observations were considered for selecting an appropriate encoding strategy for each variable. The price distribution of all the categorical levels were mostly uniform and there was not detected a significant categorical parameter which could highly affect the cars' prices.

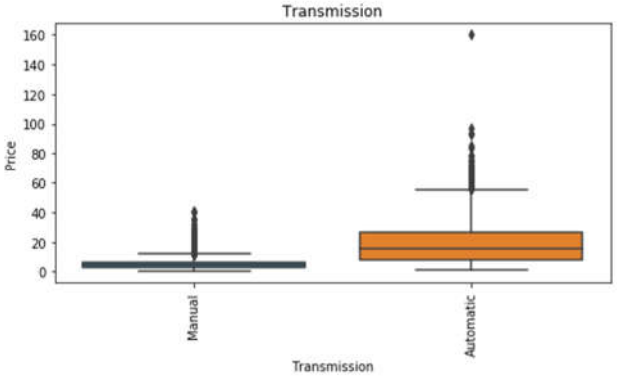

Location
count
Mumbai
Pune
Chennai
Coimbatore
Hyderabad
Jaipur
Kochi
Kolkata
Delhi
Bangalore
Ahmedabad
Location

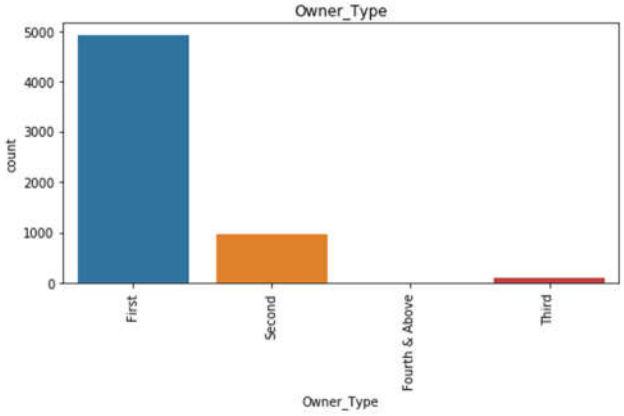

Location
Price
Mumbai
Pune
Chennai
Coimbatore
Hyderabad
Jaipur
Kochi
Kolkata
Delhi
Bangalore
Ahmedabad
Location

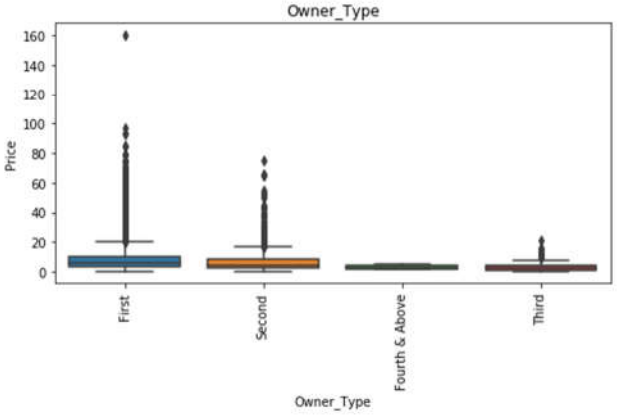

Fuel_Type
count
CNG
Diesel
Petrol
LPG
Electric
Fuel_Type


Fuel_Type
Price
CNG
Diesel
Petrol
LPG
Electric
Fuel_Type


Transmission
count
Manual
Automatic
Transmission


Transmission
Price
Manual
Automatic
Transmission


Owner_Type
count
First
Second
Fourth & Above
Third
Owner_Type


Owner_Type
Price
First
Second
Fourth & Above
Third
Owner_Type

Data Modeling

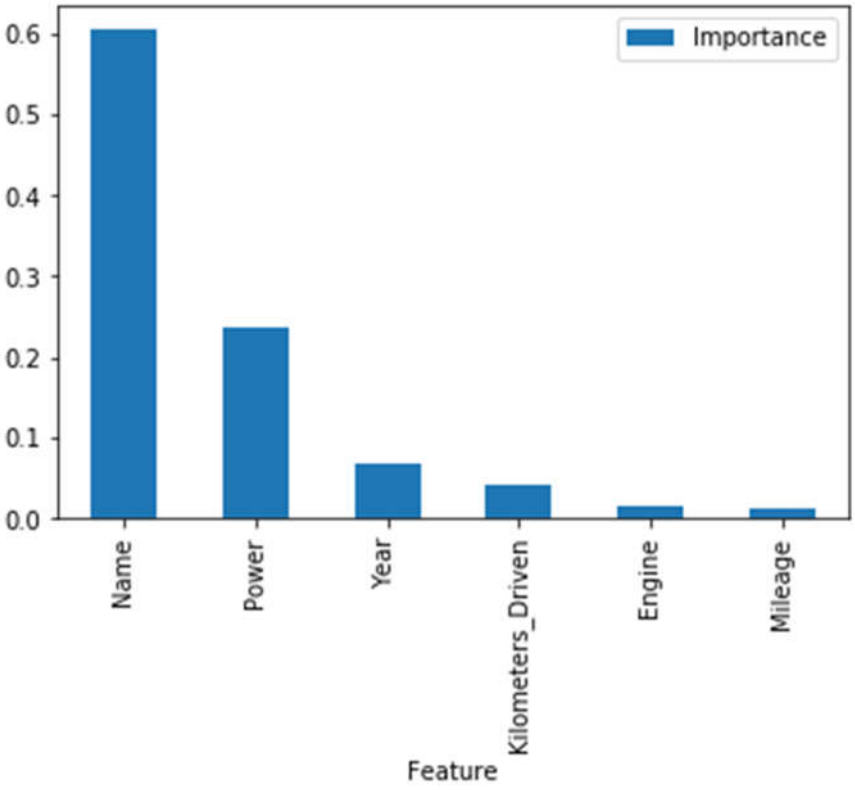
1. Feature Engineering

This phase of the project contained three steps: feature encoding, feature scaling, and feature selection. Two of the categorical variables (i.e., 'Name' and 'Location') had high number of unique values, so that we couldn't use OneHotEncoding method for them as it could cause curse of dimensionality! Therefore, I used target encoding technique to replace them with numerical values. Prior to the encoding, I merged the categorical levels with low frequencies as 'rare' incidents to reduce calculation cost. For the remained categorical variables, I replaced them with dummy variables. Feature scaling was also included in the model as an option which could be applied based on user's decision and the selected models.

2. Model Selection

Three models were selected for predicting the car price: Linear regression, random forest regression and gradient boosting regression. 'Mean squared error' was selected as the metric for evaluating models' performances. The models were built with default parameters, while 3-fold cross validation was applied for models' assessment.

The features' importances were calculated using randomforest algorithm and were shown as a table along with a barplot graph. The results showed that 'Name', 'Power', 'Year', and 'Kilometers_Driven' features were the most effective parameters in defining the used-car prices. On the other hand, 'Fuel_Type', 'Owner_Type', 'Transmission', and 'Seats' variables were the lowest important parameters in predicting the price of used-cars.



3. Model Tuning

Gradient boosting model was selected as best model for predicting the price of used cars with ‘MSE’ value of 11.8. In the next part of the project, the hyperparameters of the selected model was tuned by GridSearch algorithm which resulted in 12% improvement in the ‘MSE’ value.

The optimal model was finally trained with the whole training data, the model was used for predicting the price of cars in the test dataset, and the predictions were saved as a .csv file. I skipped the model-testing step since I didn't have the outcomes for the test data.

Conclusion

This project was an end-to-end project for building a predictive model for defining the price of used cars based on their multiple independent parameters. Different phases were performed for the data preprocessing and data exploration, before building a predictive model. The best ‘MSE’ score of around 10 was achieved after optimization of the best model.

Future Works

For future works, there are certain areas in which the modeling could be further tuned. For imputation of the missing values, more selective approaches could be utilized based on the nature of each variable. More advanced feature engineering techniques could also be applied to prepare better data for more accurate modeling.