REPORT – USED CARS PREDICTION MODEL

Contents

Executive Summary 1
Problem summary 1
Solution summary 3
Recommendations for implementation 7

Executive Summary

Create a pricing model that can accurately forecast the cost of used automobiles and assist the company in developing lucrative pricing strategies. We have to create a model that will be quite effective in resolving this issue. Regression algorithms are employed because, as opposed to categorical values, their output is a continuous value. We are coming up with a pricing model using linear regression. The data is comprehended using Exploratory Data Analysis, Univariate Analysis and Bivariate Analysis. Data cleaning is also done entails identifying all null and missing values and removing outliers. The Model with minimum Root Mean Squared Error is Linear Regression and the Model with highest accuracy is Linear regression. The models with higher adjusted and predicted R-squared values are often the ones you pick. By splitting data, cross-validation determines how well we improve decision making to other data sets.

Problem summary

Given the variety of elements that influence a used car's market pricing, determining if the quoted price is accurate, is a difficult undertaking. It is feasible to anticipate the real price of an automobile rather than just the price range of a car since regression algorithms give us a continuous value as an output rather than a categorized value. It can be challenging to decide whether a used car is worth the asking price while viewing listings online. The actual value of an automobile might vary

depending on several factors, such as mileage, make, model, year, etc. Pricing a used car fairly is a challenge from the seller's point of view. The goal is to employ machine learning algorithms to create models for forecasting used automobile prices based on existing data. According to a recent report on India's used automobile market, 4 million second-hand cars were acquired and sold in 2018–19, compared to 3.6 million new car sales. According to the IndianBlueBook research, one of the main reasons why individuals choose a used automobile over a new one is because they get good value for their money. This is hardly surprising, according to Shubh Bansal, co-founder of the used automobile marketplace Truebil, considering our country's economic growth and expanding disposable wealth. He added that the trend of rising used car sales will continue in India. In developed nations, the ratio of individuals buying a new car and a used car is 1:3, indicating that out of four people who buy a car, just one gets a new car and three choose for a used car. The used car market is a very different proposition, with significant uncertainties in both pricing and supply, in contrast to new cars, where price and supply are largely deterministic and managed by OEMs (Original Equipment Manufacturers), except for dealership level discounts, which only come into play in the final stages of the customer journey. Mileage, brand, model, year, and other elements can all affect a car's true value. Setting the right price for a used car is not a simple process from the seller's standpoint. The pricing structure of these secondhand cars becomes crucial for market expansion.

We have to create a pricing model that can accurately forecast the cost of used automobiles and assist the company in developing lucrative pricing strategies. We must create a model that will be quite effective in resolving this issue. Regression algorithms are employed because, as opposed to categorical values, their output is a continuous value.

The Crucial Issues include Do different influencing elements have an impact on a used car's price? What are the numerous independent factors that affect the price of used cars? Does the cost of an automobile depend on its name? How does pricing depend on the type of transmission? Does the price of the car depend on where it is located? Does the price of a car correlate with the number of miles travelled or the year it was made? Does the price of a car have anything to do with its engine, power, or mileage? What impact do seat count and fuel type have on price?

Solution summary

There are a total of 7253 rows and 13 columns. The 13 columns are - S.No. : Serial Number , Name: Name of the car which includes Brand name and Model name , Location: The location in which the car is being sold or is available for purchase (Cities) , Year: Manufacturing year of the car, Kilometers_driven: The total kilometers driven in the car by the previous owner(s) in KM , Fuel_Type: The type of fuel used by the car (Petrol, Diesel, Electric, CNG, LPG) , Transmission: The type of transmission used by the car (Automatic / Manual), Owner: Type of ownership, Mileage: The standard mileage offered by the car company in KMPL or KM/KG, Engine: The displacement volume of the engine in CC, Power: The maximum power of the engine in BHP, Seats: The number of seats in the car, New_Price: The price of a new car of the same model in INR 100,000, Price: The price of the used car in INR 100,000. S.No,Name,Location,Year,Kilometers_Driven,Fuel_Type,Transmission and Owner_Type have no missing values. Mileage , Engine, Power , Seats, New_price and Price have missing values. Name, Location, Fuel_Type, Transmission, Owner_Type are of data type object and the rest of the columns are of numerical data types. We can observe that S.No. has no null values. Also, the number of unique values are equal to the number of observations. So, S.No. looks like an index for the data entry and such a column would not be useful in providing any predictive power for our analysis. Hence, it can be dropped. The average year is 2013 with the oldest car from 1996 and the newest car is from 2019. Kilometers driven has an extremely high standard deviation as the value is greater than the mean by a lot. Many columns have empty values so the data cannot be accurately summarized. From the summary , average number of cars have 5 seats. 50% of the cars have a mileage of 18.16 or lesser. 75% of the cars have the original price of 995000 and a new price of 260425. There are approximately 2041 unique values in the column 'Name' , 11 unique values in 'Location' , 5 unique values in 'Fuel_Type' , 2 unique values in 'Transmission' and 4 unique values in 'Owner_type'. The maximum value of Kilometers driven is 65000000 with a price of 650000 and the minimum value of kilometers driven is 299322 with a price of 400000. Kilometers_Driven is highly right-skewed , so we us log transformation to reduce or remove the skewness. Just like Kilometers_Driven, the distribution of price is also highly skewed, so log distribution is done on this column to see if the data can be normally distributed. New_price is positively skewed to the left. The outliers for New_price is more than 532.787.5. Year has only few outliers on the left and is slightly negatively skewed. Mileage is positively skewed has one

outlier in the left and three in the right. Engine is positively skewed with outliers greater than 3123. Power is positively skewed with outliers greater than 32.75. 13.1% of cars are from Mumbai, 12.1% of cars are from Hyderabad. Coimbatore, Kochi and Pune have approximately 10.6% cars each. Approximately 9% of cars are from Delhi and Kolkata each. 8.1% of cars are from Chennai, 6.9% of cars are from Jaipur , 6.1% of cars are from Bangalore , 3.8% of cars are from Ahmedabad. %3.1 % of cars use Diesel , 45.8% use Petrol and 1.1% of cars use CNG or CPG or Electric. 71.8% of cars are Manual Transmission and 28.2% of cars are automatic. 82.1% of cars have Owner types as First, 15.9% as second, 1.9% as third and 0.2% of cars are fourth and above. As the year increases , the price of the car increases. Price has no correlation to the kilometers driven. The Price increases if the Power and Engine increases. Price has no correlation with Seats and Mileage. Engine and Mileage have a very strong negative correlation. The vehicle's brand can be found, and the remaining details can be ignored. Should substitute a "brand" column for this one to reflect the brand of the car. The goal variable is positively skewed, as seen by the histogram, thus, to improve the linear models for regression, we must log-transform this data to make it behave more normally. The boxplot demonstrates how noisy this target variable is. Should remove all outliers of columns so that the regression model can be accurately formed.

Splitting the train-test data should be done. In this instance, we separate the data into training and test sets, fit potential models on the training set, assess them on the test set, and choose them. Check for multicollinearity using the Variance Inflation Factor (VIF) , Features having a VIF score > 5 will be dropped / treated till all the features have a VIF score < 5. Since it is impossible to predict, we are unable to predict which model will solve this problem the best. As a result, we fit and assess a variety of models to the issue. On the train set, we apply Linear Regression Models (Feature Selection). The process of developing machine learning models must include evaluating the model accuracy in order to characterize how well the model is performing in its predictions. In regression analysis, the MSE, MAE, and RMSE metrics are primarily used to assess model performance and prediction error rates. The difference between the original and anticipated values, as determined by averaging the absolute difference over the data set, is represented as MAE (mean absolute error). The difference between the original and predicted values is represented by the MSE (Mean Squared Error), which is obtained by squaring the average difference throughout the data set. The error rate by the square root of MSE is known as RMSE (Root Mean Squared Error). By applying a linear equation to the observed data, the linear regression method attempts to model

the relationship between two variables. The second party is regarded as the dependent variable. Finding a relationship between several continuous variables can be done with the help of linear regression. Both several independent variables and a single independent variable are present.

According to the coefficients and intercept in the model, Transmission_Manual, Location_Kolkata ,Kilometeres_driven_log,Fuel_type_LPG,Fuel_Type_Petrol,Owner_type,Locatton_Pune,Locatio n_Mumbai,Location_Jaipur, Location_Chennai ,Location_Delhi has a negative coefficient which means that it has a negative correlation with the dependent variable. Engine,Power,Mileage,Location_Hyderabad,Location_Bangalore,Seats,Location_Kochi,Fuel_T ype_Dieasel,Location_Coimbatore have a positive coefficient, so they are positively correlated. In the Linear regression model , the mean absolute error, which is the average of the absolute difference between the actual and predicted values in the dataset is 0.60,the root mean squared value is around 0.8, the r- squared value is 0.4 , and the adjusted r-squared value is 0.4. Through the calculation of the mean of residuals, the regression line passes through the points. The p-value is not less than 0.05,so we cannot reject the null hypothesis, therefore the data is homoscedastic. The residual terms are normally distributed. According to the model, the following factors are the most important determinants of used automobile prices: year, number of seats, engine power, mileage, kilometers driven, owner type, fuel type, location, automatic/manual transmission, and new-car price.

The coefficients can help with the interpretability of the model. Looking at the coefficients Transmission_Manual affects the new price by approximately -26% , Location_Kolkata affects the new price by approximately -23% , Fuel_Type_Petrol affects the new price by approximately -17.5 % , Location_Pune affects the new price by -9.8% , kilometers_driven_log affects the new price by -8 %, Fuel_Type_LPG affects the new price by -6.7% , Location_Jaipur affects the new price by -5.6% , Location_Kochi affects the new price by -5.5% , Location_Mumbai affects the new price by -3.5% , Location_Chennai affects the new price by -1.7% , Location_Delhi affects the new price by -1.2% , Mileage affects the new price by -1.2% , Owner_Type affects the price by -1% , Seats affect the new price by -0.7% . Fuel_Type_Electric and Engine do not have that much affect on the new price. Power affects the price by 0.5% , Location_Coimbatore affects the price by 2.4% , Location_Bangalore affects the price by 3.2% , Location_Hyderabad affecs the price 7.8% , Fuel_Type_Diesel affects by 9.1% and Year by 9.5%.

We convert the categorical variables to numeric format as most of the algorithms cannot handle non-numeric data. I used One-hot encoding. One-hot Encoding is only use binary values (0s, 1s) to represent category data. If the categorical feature has few unique values, one-hot encoding is preferred over label encoding. The elements that have the biggest impact on "Price" will be chosen next. With the exception of gender, which has very little impact on "Price" I have chosen all of the attributes. The 'x' variable will be these features, while the 'y' variable will be charges. Regression model evaluation measures include mean absolute error (MAE) and root-mean-square error (RMSE). Let's adjust a few algorithmic parameters and evaluate the model's precision. It takes a lot of time to manually test out various parameter value combinations. This procedure is automated by Scikit- learn's GridSearchCV, which also determines the best values for these parameters. The distribution and residual plots show that the expected and real charges have a considerable amount of overlap. However, a few of the projected values are significantly off the x-axis, which increases our RMSE. By expanding our data points, or gathering additional data, we can lower this.

Using feature importance in decision tree regressor , we can see that the top 5 variables that are important for calculating price. Engine is around 6-7% important, Mileage is around 8% important , Year is 14% important , Kilometers_Driven is around 21% important , Power is around 35% important.

The RMSE is a useful indicator of prediction accuracy since it shows the average amount of mistakes in the projected data. The R2 measure, where x and y represent a collection of data, displays the percentage of variation in y that is explained by x-variables. In the Linear Regression Model , the Mean Squared Error , Root Mean Squared Error , Explained Variance Score and R-Square Score/Accuracy is 0.623832 , 0.79267, 0.43085 and 0.43062 respectively. In Decision Tree Regressor Model , the Mean Squared Error , Root Mean Squared Error , Explained Variance Score and R-Square Score/Accuracy is 1.27686 , 1.12998 , -0.15520 , -0.15709 respectively. Random Forest Regressor Model has the Mean Squared Error as 0.66025 , Root Mean Squared Error is 0.81256 , Explained Variance Score is 0.40319 and R-Square Score/ Accuracy is 0.40168. In Ridge and Lasso model , the Mean Squared Error is 0.62835 , Root Mean Squared Error is 0.62835 , Root Mean Squared Error is 0.79268 , Explained Variance Score is 0.43082 and R-Square Score/Accuracy is 0.43059.

Recommendations for implementation

An understandable statistical scale known as the R2 Regression Score or Coefficient of Determinant assesses the percentage of a dependent variable's change that is explained by one or more independent variables in a regression model. Constant models always predict the expected value of y regardless of the input features. For a model that performs arbitrarily poorly, R2 value can be negative. R2 is usually used to gauge how well a model fits the data. However, this metric has a weakness that should be mentioned since the R2 score always tends to rise with more features without necessarily enhancing the model's fit. A measure of how closely a regression line resembles a set of points is called the Mean Square Errors (MSE). To eliminate any negative indications, it squares the residuals' errors, also known as their variance, which are the distances between the points and the regression line. The Root Mean Squared Errors (RMSE), which is calculated by taking the square root of the MSE, shows how well the model fits the data in terms of how closely the observed data points match the values predicted by the model. Considering that RMSE has the same unit as the response variable, it is significantly simpler to read. A better fit is indicated by lower values of RMSE, which indicate that the regression line is near the data points. The models with higher adjusted and predicted R-squared values are often the ones you pick. These statistics are intended to overcome a significant issue with standard R-squared, which is that it grows with each additional predictor and may lead you to build an unnecessarily complicated model. The adjusted R squared can go down with bad predictors and only goes up if the additional term enhances the model more than would be predicted by chance. A type of cross-validation is the projected R-squared, which can also fall. By splitting data, cross-validation determines how well we improve decision making to other data sets. A low RMSE indicates that the residuals are close to zero in terms of the magnitude of the response variable. So Linear regression model would be the best option for predicting the price followed by ridge and lasso regression , random forest regressor and decision tree. For interpretability of the model , the coefficients from the regression model and the features importance would be beneficial.

The final Linear regression model has a MAPE of 35-50% which means we can predict the price within that range of the price value. It might be wise to stock newer vehicles because they will be more profitable because they produce higher pricing and have fewer owners. Powerful automobiles

with diesel engines and automated transmissions frequently fetch the highest prices. For the best return, look for these vehicles. Vehicles with lower usage will cost more. Purchase automobiles in markets with cheaper prices (Kolkata, Delhi, and Mumbai) and resell them in markets with higher prices (Bangalore, Hyderabad, Coimbatore).