# Price and Symbolling Prediction Model of an Automobile

## Non-Technical Summary

**Harsh Shukla**
**Khushru Irani**
**Nisarg Patel**
**Rama Mani Deepika Maram**
**Trupti Kirve**
**Vivek Bhavsar**

In our below research activity, we are trying to investigate an application based on the supervised machine learning techniques to predict the price of the vehicle. Determining the price of any commodity is typically an interesting problem which has the continuous predicted variable.

With the increasing number of choices available in the market from several brands, different types of vehicle engines, the performance of the vehicle makes this research more interesting and hence such predictions help the common man to buy a vehicle.

It is daunting when we think of buying a vehicle even for personal use. The number of options available for several brands. The engine type, size. The performance of the vehicle everything matters. You can think of how difficult it is to make such a decision which has been based on many such parameters. Hence, we think that our research of predicting the price of the vehicle will give a little relief to everyone who is looking to buy a vehicle.

The predictions are based on data collected from UCI Machine Learning library. The data set consist of 205 instances with 26 variables (including price). Our project work, moves around these variables to identify how they influence the price of the vehicle. The data was collected as a part of 1985 Ward's Automotive Yearbook [1], Personal auto manuals of Insurance Service Office [2] and Insurance Collision Report by the Insurance Institute for Highway Safety [3]. Ward is an American organization which has covered the automotive industry for about 80 years.

Fortunately or unfortunately, the selected dataset did not have a lot of missing values. The instances where the value was missing for the dependent variable which summed up to 4 were removed from the dataset. For the rest of the independent variables like normalized_losses we filled in values by considering the mean of the nearby points. For variables like no_of_doors, bore, stroke, horsepower and peak_rpm we filled in values manually by doing some research based on make, engine, type of automobile etc.

Exploratory analysis using various techniques of visualization is carried out to bring the important aspects of the data into focus for further analysis. Summary statistics can be used to find tendencies of the data and its quality to formulate assumptions. Descriptive statistics on numeric metrics is carried out to understand the mean, standard deviation etc.

1. 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.
2. Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038
3. Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

The selected dataset shows high correlation and if we come up with a model using the dataset directly we will face multicollinearity and high Variance Inflation Factor. The issues with multicollinearity and VIF then in-turn raise the issue of Overfitting. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

To avoid the above issues, we performed initial exploratory techniques on the selected dataset. The techniques include Principal Component Analysis, Common Factor Analysis, Canonical Correlation Analysis, Correspondence Analysis. These techniques explain the percentage of the variance explained by the model.

To come up with prediction model we used two machine learning methods namely Linear Regression Model and Decision Tree. Linear Regression Model is used to predict the price of the automobile using the various features available, whereas we created a decision tree model to predict the symbolling features (symbolling is the risk associated with the automobile depending on various another feature).

The linear Regression Model used shows a high R-squared value meaning it explains high variance in the data and hence we can suggest this model to predict the price of the automobile. Also, the model involves few features this helps achieving parsimony.

The decision tree is used for determining the level of risk in car insurance in which the symbolic variable was used as target to build the model. Moreover, the results gave high accuracy for both the training and testing set.

As both the models have performed analysis for different features the models are not comparable, and we say use Linear Regression Model when we have to predict the price and when we need to identify the risk associated with the vehicle we can use the Decision Tree model.

After performing all the above analysis on the selected dataset, we also plan to perform Linear Discriminant Analysis on the categorical variables and K-means Clustering technique to identify the groupings of the instances.