PREDICTIVE ANALYTICS COMPETITION (PAC) REPORT

1. Introduction

The used cars competition represents an exciting open-ended challenge in the realm of analytics. The primary aim of this competition was to develop predictive models capable of predicting the price of used cars accurately, based on a wide variety of features. This report presents a detailed analysis of the strategies, methodologies, learnings and results obtained throughout the competition.

2. Data Summary

The data provided for this challenge consists of 45 factors associated with a car along with the predictor variable “*price*” for the used car. Out of these factors, 18 are numericals, 12 are categoricals, 8 are boolean variables and the rest are of mixed data-types. There are 40,000 samples in the analysis dataset and 10,000 samples in the test dataset. Many of these columns contain NA values and blanks in them, here’s the summary of the columns with the highest occurrences of missing values.

Column	Percentage NA’s
owner_count	49.39%
highway_fuel_economy	10.81%
city_fuel_economy	10.81%
mileage	4.5%
seller_rating	1.35%
engine_displacement	0.7875%
horsepower	0.7875%
fuel_tank_volume_gallons	0.0075%
exterior_color	0.0075%

Column	Percentage Blanks
is_cpo	94.06%
fleet	46.95%
frame_damaged	46.95%
has_accidents	46.95%
isCab	46.95%
salvage	46.95%
franchise_make	24.27%
torque	12.475%
power	11.0475%
major_options	5.34%

Submitted by Rihaan Satia rs4496

3. Data Exploration and Cleaning

Data Exploration Performed:

- Data Type of Columns and Conversion to correct data type
- Correlation Tests
- Evaluating mean values amongst different groups of categorical variables
- VIF

Observations:

- *wheel _system* and *wheel_system_display* show multicollinearity
- *wheelbase_inches* and *length_inches* have high correlation, hence using only one of them might improve predictability
- *front_legroom_inches* and *back_legroom_inches* are not very correlated, hence they can be used together

Imputation:
From the variables stated in the table, the ones with a high number of NA values were imputed with “unavailable”
For the rest of the variables, various imputation methods were tried. Finally, Predictive Mean Matching was used for numerical variables and Polynomial Regression was used for categorical variables.

4. Feature Engineering

Feature engineering is a crucial aspect of developing good predictive models. By crafting relevant and informative features, we can provide the model with clearer and more focused information, helping it learn patterns and relationships more accurately. Carefully designed features can make the model's reasoning more transparent and understandable, providing valuable insights into how it reaches its decisions.

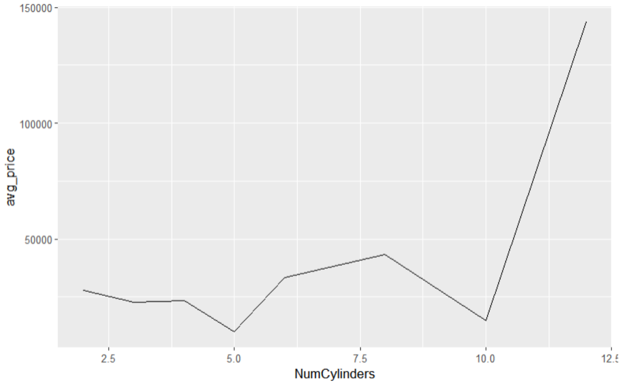
Feature engineering plays a central role in this model, with specific features extracted as follows:

- Number of cylinders*:* Extracted from the engine type. For example, V8 means 8 Cylinders
- Engine Initial: Initial of the engine type For Ex. H, V, I etc.
- Flex fuel: Indicates whether the engine is flexible fuel-type
- Transmission number: Extracts the gear number from the transmission display. For example, 6 is extracted from 6-speed Automatic
- RPM: Rotations per minute extracted from power. For example, 5600 RPM from a power of 170 hp @ 5,600 RPM
- Interior color: Initial color extracted from interior colors. For example, black from Black (Ebony)

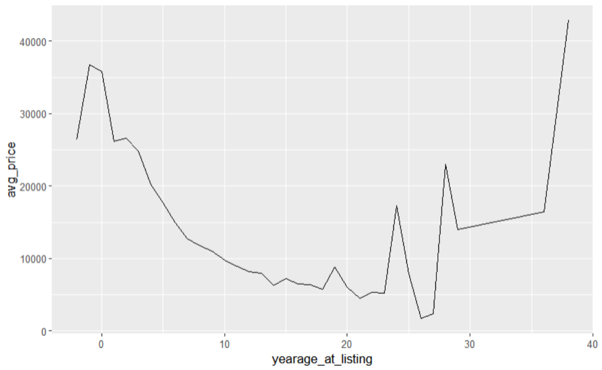
- Year-age at listing: extracted by subtracting current date from listed date. Gives the number of years since the car has been listed
- Numerous columns extracted from major options which pertain to car features, all of these are booleans:
  Roofs(Sunroof/Moonroof), third row, heated seats, leather, cruise control, blind spot monitoring, Premium, climate control, luxury, sport, steel wheels
  TRUE: Car possesses said feature
  FALSE: Car doesn't possess said feature
- Size: Obtained by multiplying the height, width and length of the car
- Average economy: Obtained by taking the average of highway economy and city economy
- Brand Type: Obtained by categorizing the brand of the car into the following distinct groups:
  Very Nominal, Nominal, Average, Above Average, Luxury, Very Luxury, Ultra Luxury

Apart from this, I also tried implementing Principal Component Analysis. I did so by generating 12 Principal Components for my entire predictor variable set. However, while trying to fit models, it didn't seem to work very well

5. Visualizing Relationships



Average Price vs Number of Cylinders



Average Price vs Year Age at Listing

6. Model Development

After completing all of data cleaning and feature engineering, various models were trained and tested using the imputed datasets. Analysis data was split into train and test using a ratio of 0.7 and the *createDataPartition* from the *caret* framework

MODEL	TRAIN RMSE	TEST RMSE
Linear Regression	9149.121	8921.763
Decision Tree	9553.551	9459.869
Decision Tree with all predictors	8109.673	7874.885
Bagging	9024.989	8855.971
Forest Ranger	2003.532	3886.772
Forest Ranger Tuned	1709.113	3701.912
XG Boost	1123.731	2137.271

I believe I could have tried selecting various combinations of features for my best models i.e tuned Ranger and tuned XG Boost. Selecting the right set of variables after extracting numerous features from the data could have helped in getting the RMSE lower. Also, I was not able to perform end-to-end hyper-parameter optimization. Due to the large number of predictor variables after performing dummy coding, having a larger tuning grid was computationally not feasible. Reducing predictor variables could have made space for better parameter optimization.

I finally achieved the lowest RMSE Score of 2292 later (with some variable changes), and my lowest RMSE in Kaggle during the competition was 2137, which was achieved by an XGBoost model.

7. Learnings and Conclusion

A problem like predicting the price of a car is quite open-ended. The analytics competition has been a journey of discovery and learning, offering variable insights into the do’s and don'ts of predictive analytics. One of the most important aspects learnt through this competition is Data Exploration and Understanding. A machine learning model needs its predictors to be in a form where it can extract predictive power out of them, and that form can only be achieved by applying the right data transformation techniques. Lastly, even small parameter changes go a long way in terms of predicting, and hence, tuning the model to its best version is crucial.

NOTE: Please find attached RMD with detailed code.