Sirak Jraghatspanyan

Final Report
Predicting the “car kicks”

Problem Statement

One of the biggest challenges of an auto dealership purchasing a used car at an auto auction is the risk that the vehicle might have serious issues that prevent it from being sold to customers. The auto community calls these unfortunate purchases "kicks". Kicked cars often result when there are tampered odometers, mechanical issues the dealer is not able to address, issues with getting the vehicle title from the seller, or some other unforeseen problem. Kick cars can be very costly to dealers after transportation cost, throw-away repair work, and market losses in reselling the vehicle.

By using the 2009-2010 US auction sales dataset I created a tool that can help to find which cars have a higher risk of being a kick. This tool can provide real value to dealerships trying to provide the best inventory selection possible to their customers.

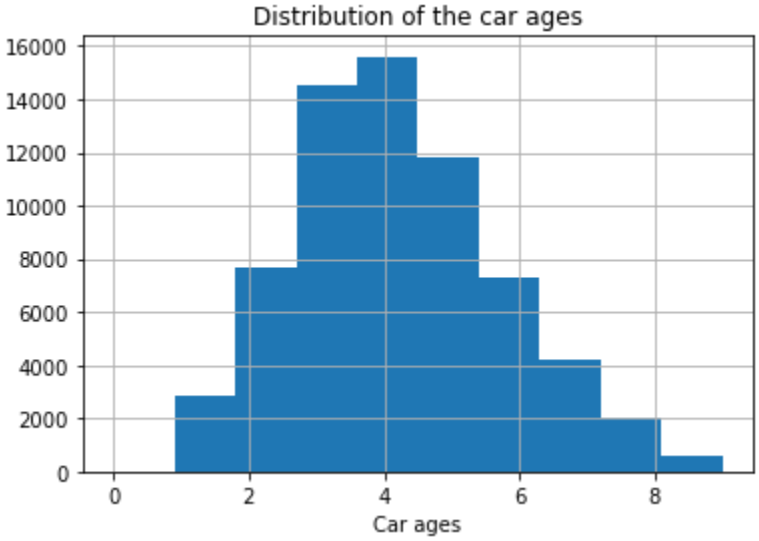
Data Wrangling

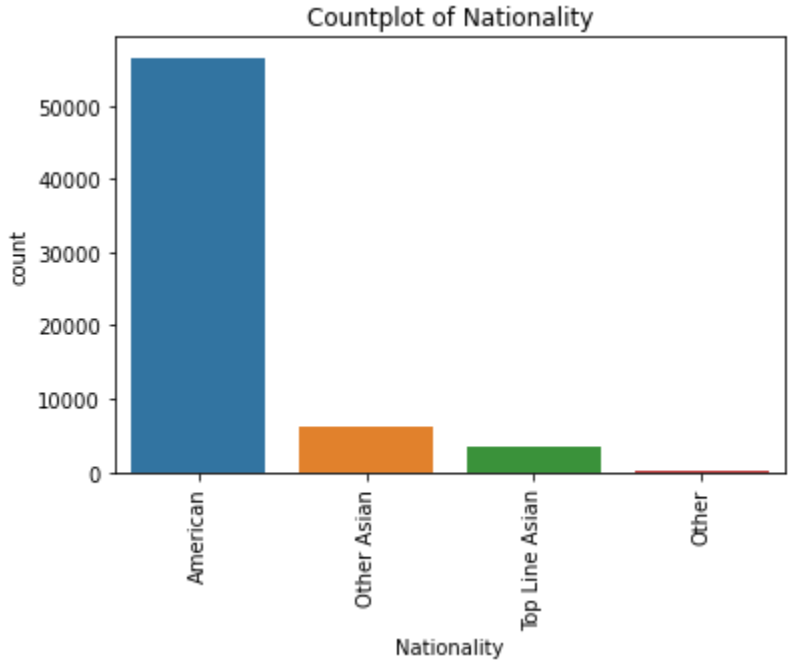
The original dataset had 67211 rows with 29 columns. It didn’t have any missing values but it did have 2596 rows with 0 price values which were eventually dropped. The data types were changed for 4 of the columns.

The ‘Model’ column representing the car model was cleaned by keeping only the first two words of it containing not less than 4 characters which decreased the number of unique values of the column 3 times. Some of the columns contained quotation marks which were removed. I also found 2 pairs of duplicate rows which were dropped.

Exploratory Data Analysis

As I mentioned the dataset contained information about auctions between 2009 and 2010.



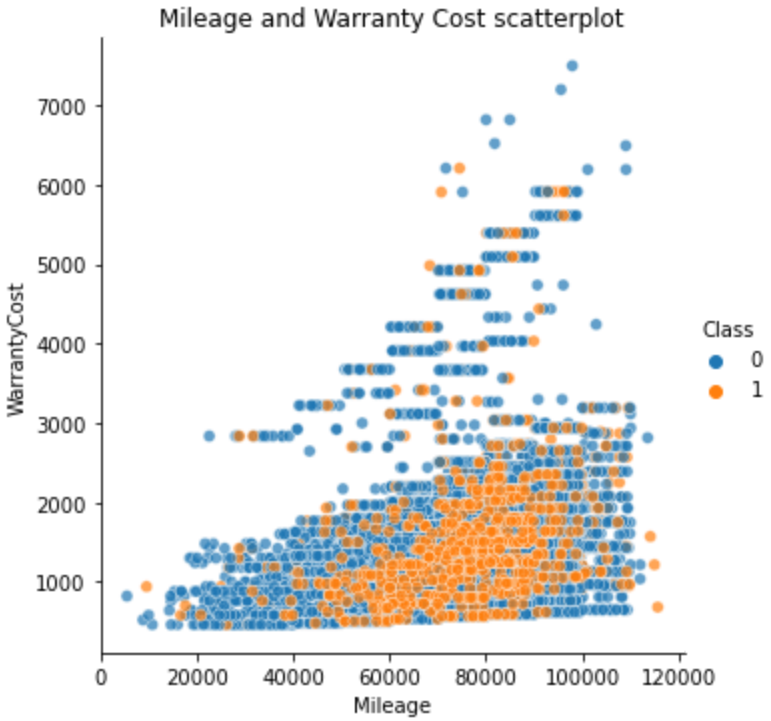


Next we had columns representing the year of production and age of the cars: from the graph on the left side we can see that most of the cars were 2 to 6 years old. Most of the cars have mileage between 60.000 and 90.000 which are average values for 3-6 years cars. The price columns have similar distribution and are close to normal.

And regarding the categorical features it appeared that most of the cars sold on these auctions were medium size american cars.

Also we tried to find out how different features were correlated with each other. There is some strong correlation between the year and price columns.

And we can notice some moderate correlation between the mileage and warranty cost, and between these columns and age of the car.



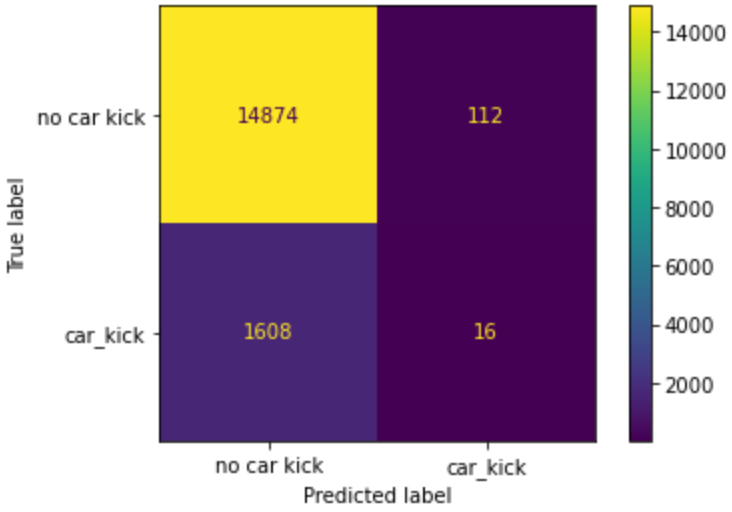
We can see the moderate positive correlation between the warranty cost and mileage. An interesting insight from this graph was the fact that most of the car kicks have mileage between 40.000 and 90.000. And considering the fact that one of the main criterions of being a 'kick' car is the rolled back odometer it can be the case here.

Preprocessing and Modeling

In the preprocessing step some columns were dropped and by the end of this part the dataset had 16 features. The categorical columns were transformed into dummy variables. After the features were scaled and splitted into testing and training datasets.

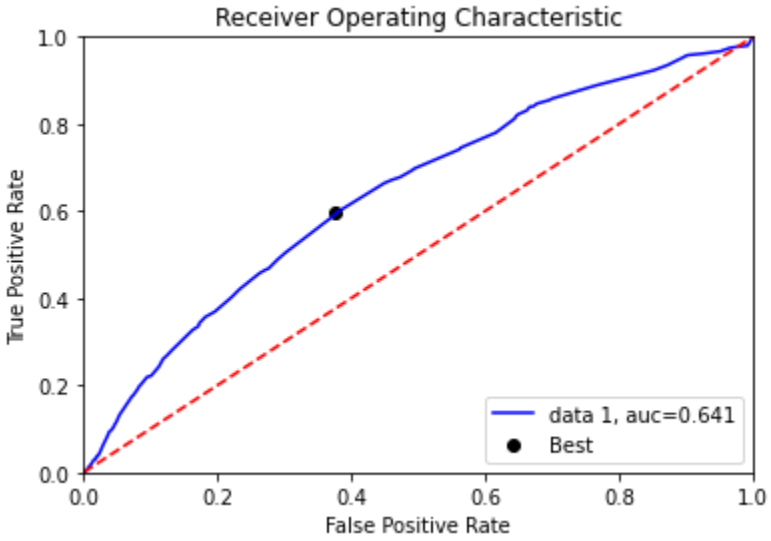
For the modeling step I tested 2 machine learning classification models: Logistic Regression and Decision Trees. Logistic Regression with the default parameters performed poorly on our data. Next we tried the ‘saga’ algorithm for the solver parameter in Logistic Regression but it didn’t improve the model. The ROC AUC score was equal to 0.5.

For the Decision Trees models we tried it first with max depth of 5 which didn’t perform well. Next we tried to tune our hyperparameters changing max depth and criterion using Grid Search. It showed that the best parameters for the Decision Tree model are criterion: 'gini' and max depth of 10. This model has a ROC AUC score of 0.64.



But the confusion matrix showed that even though the roc auc score is higher than in case of the previous models, the base threshold value of 0.5 doesn't work well for car kicks: only 1% of the car kick cases were predicted right. I needed to find the optimal threshold where the recall value for car kicks was the highest.

I found that the best threshold is 0.0909 for our model where the recall is 0.5954, which is the black dot on the graph below. So 59% of the car kick values were predicted right.



Takeaways

So the Decision Trees model with max depth of 10 and gini criterion was the best fit for the dataset. But it still didn't perform well on car kicks. Finding the optimal threshold improved the performance on car kicks but it would be hard to apply this model on practical data with the recall value of 0.59. One of the main reasons of this result is the highly imbalanced dataset which makes it harder for any model to perform and also the features

in the dataset were not the best ones for identifying the car kicks considering the fact that being a car kick is more connected with the mechanical details of the car rather than its appearance and manufacturer information.

But the model still can be useful for used car dealers in the process of identifying the car kicks by providing some general and easily accessible information about the cars they are interested in.

Future Research

As I mentioned above the model can be improved by adding more mechanical information about the cars in the data. For example, were there any recalls on that model of the car, what is the reliability rating for it, how much is the likelihood for the model to be in a car accident. Combining these features with the more balanced dataset can significantly improve the model's performance.