Observing the Effects of the Curse of Dimensionality

Nadiem Ahmed

Problem Statement

- Accurately predict the time it takes for Mercedes to test a car given a set of custom features that the car will have
- Observe the effects of the curse of dimensionality on the dataset/model performance

Hypothesis

- The data suffers from the curse of dimensionality
- Reducing the dimensionality of our data will improve performance

Background Information

Supervised Learning:

- A subset of machine learning where we are given features and labels, and we create a model that to map the function between feature and labels

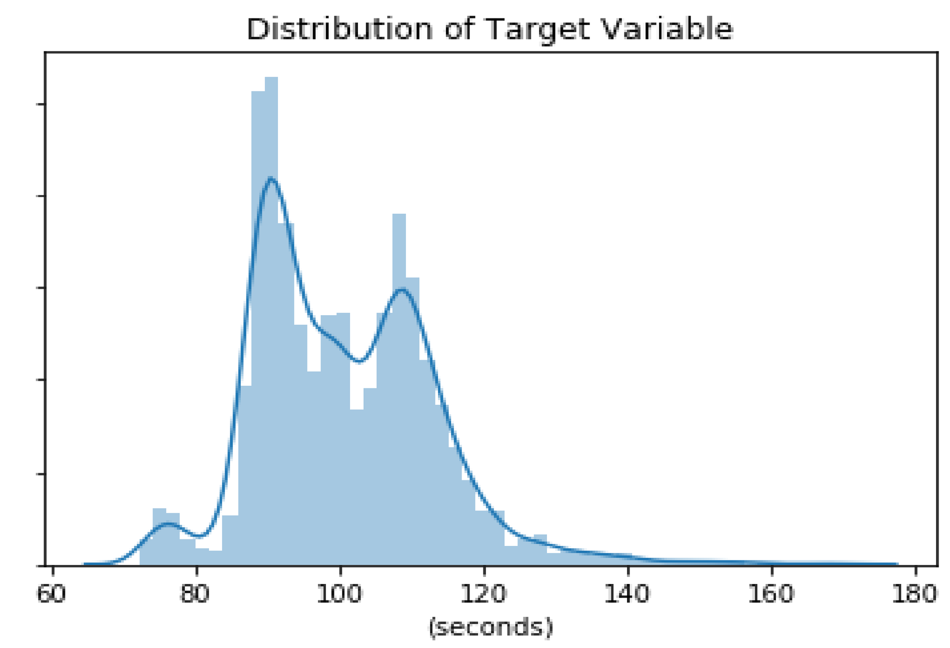
Curse of dimensionality:

- The number of samples required increases exponentially with the number of features

Coefficient of multiple determination (R^2):

- the proportion of variation in the dependent (target) variable that can be predicted from the set of independent variables (features)

Dataset



- Used a histogram to visualize the target variable
- We can see that almost all cars pass testing in 75-125 seconds
- 377 features: 8 categorical, 369 binary
- 8416 total samples-4208 in each train and test sets

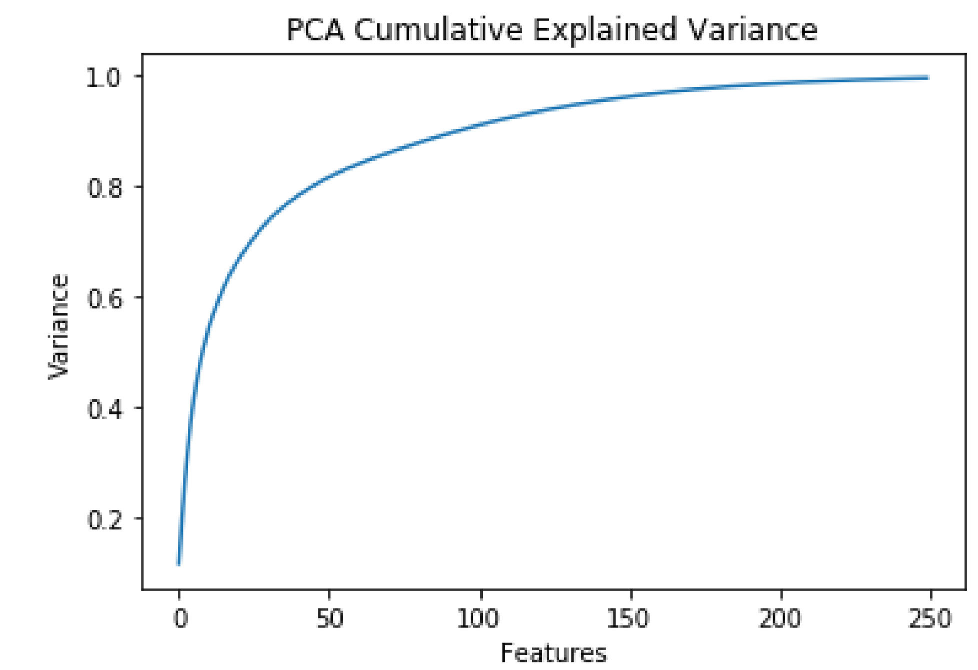
Preprocessing

- Removed columns that had only one unique value- does not add any useful information
- Used one-hot encoding to convert categorical features into binary → 541 features

Regression Models

Regression Model	Hyperparameters	R^2
Lasso	alpha=0.25, max_iter=100	0.54567
KNN	neighbors=30, leaf_size=30,	0.48104
Random Forest	max_depth=2, n_estimators=100	0.48234

Results



- From this graph we can see that around 150 features captures approximately 95% of the variance
- Used PCA to reduce dimensionality to 138 principal components, which is 95% of the variance

Model	R^2
Lasso before PCA	0.54567
Lasso after PCA	0.53616

Conclusion

The loss of information from reducing the dimensionality outweighs the effects of the curse of dimensionality for this dataset