Nadiem Ahmed November 18,2019

Project Deliverable 3

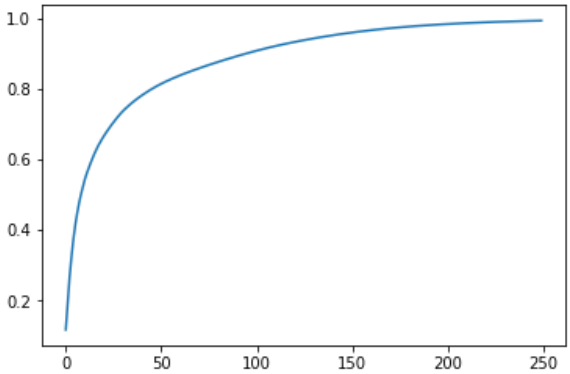
1. **Final Training Results**

R^2 value for different regression models:

Before reducing the dimensionality, I wanted to experiment with different regression algorithms. The three algorithms I decided to use were Lasso regression, Random forest regression, and KNN regression. My initial model was a Lasso regression model, and it coincidentally ended up performing the best.

Algorithm	weights/hyperparameters	R^2
Lasso	alpha = 0.025	0.54567
KNN	neighbors = 20	0.48234
Random Forest	max_depth=2, random_state=0, n_estimators=100	0.48108

After trying out different regression models, I wanted to see what effect reducing the dimensionality would have on the performance of my model. I hypothesized that the data I was working with was suffering from the curse of dimensionality, and therefore that reducing the number of features would improve performance.



I decided to use PCA. From this graph, we can see that about 150 features account for around 95% of the variance of our entire dataset, which has over 500 features. I then used PCA with n_components=0.95, which resulted in a dataset of 138 features. I then recreated the Lasso

regression model with this reduced data set. To my surprise, the performance of this model was worse than the initial Lasso regression model. It does make sense though, as we are losing information by doing PCA.

2. Final Demonstration Proposal

Initial Question: The initial problem is to create a model that predicts the time it takes for a Mercedes car to pass testing given a set of custom features the car will have.

Research: My research included looking at others notebooks' on Kaggle, looking up regression methods, and learning about the curse of dimensionality in class.

Hypothesis: I hypothesized that the data I was working with suffered from the curse of dimensionality, and that reducing the number of features would improve performance.

Experiment: To test the effects of dimensionality reduction, I had to test a single model before and after reducing the dimensionality. Before I could do that, I had to preprocess and visualize the data. Next, I experimented with 3 different regression models on the initial data set, and found that Lasso regression performed the best. I then performed PCA to reduce dimensionality, before trying the Lasso regression model on that data set.

Data/Analysis: The R^2 value of our initial model was higher than our model on the PCA data.

Conclusion: I reject my initial hypothesis. I believe that the loss of information causes a decrease in performance.