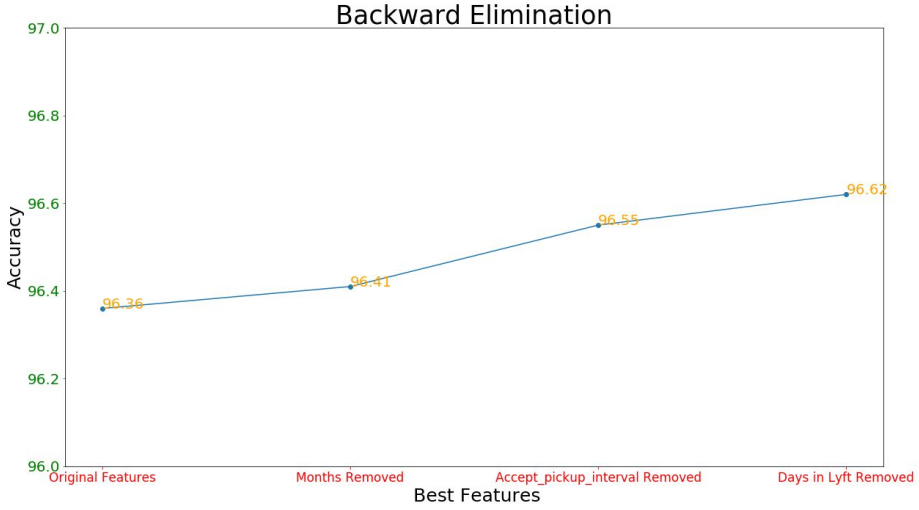
Lyft Data Challenge

Team ikun

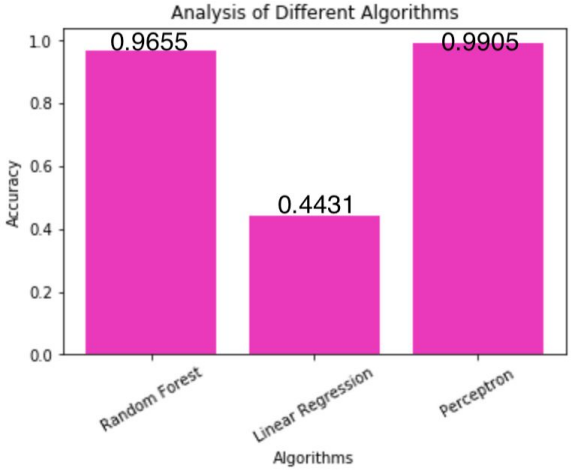
Zhen Jiang & Yihua Xu

Conclusion

With the data set provided, the average lifetime value for each driver in Lyft is around **$2979.856**. In order to maximize the lifetime value for Lyft, we examine the potential factors that contribute to the lifetime value. After analyzing the contributions each factor impacts, we found four features that really affect the outcome: Days spent with Lyft, the percentage that each driver works during prime time, total distance within his career, and total time serving as a Lyft driver.



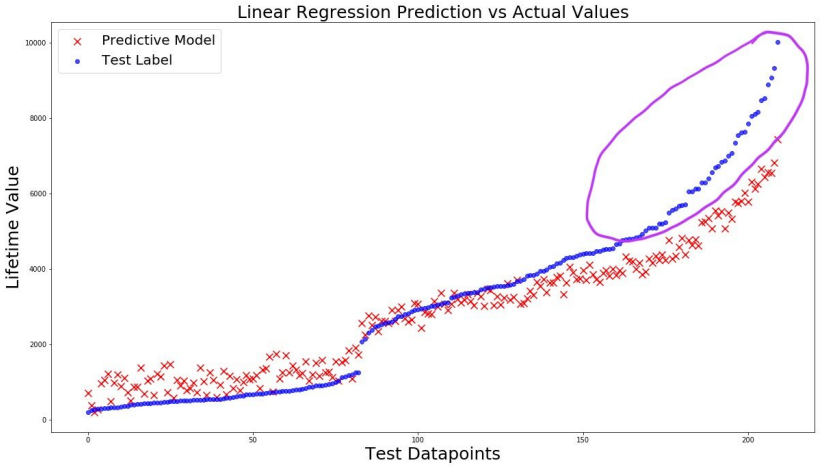
By utilizing the "random forests", "linear regression" and "perceptron"(classifier), we calculated the MAPE (Mean Absolute Percent Error). The best estimation(random forests) we yield gives us **3.450%** of error, which means that we can estimate the lifetime value with the accuracy of **96.550%**.



Graph 1
(the accuracy shown in Graph 1 is the value from 1 - MAPE)

Meantime, we also calculated the average projected lifetime of a driver. The average days they spent with Lyft are approximately **54.184.** As expected, not all drivers act alike.

What we also found is that the lifetime value is more like an exponentially increasing curve(as graph 2 suggests), rather than a linear line with respect to our four main selected features.



Graph 2

Recommendation:

1. Based on our findings, try to extend the overall lifetime for a driver in Lyft. (As the graph 2 suggests, the purple circle gives Lyft much higher value)
2. Encourage drivers to work more often during prime time because it boosts revenue easily.

Methodology by steps

I. Clean datasets, in regard to ride_ids.csv, we can find multiple records that do not match any driver in the driver_ids.csv. Also, we found several entries in ride_timestamps.csv that can not point to any ride recorded in ride_ids.csv. Consequently, we deleted all of these unusable data.

II. Prepare for the future calculation by assign each ride to a specific driver. Then calculate each driver's lifetime value by adding up every fare corresponding to each trip.

III. Select features that contribute to the lifetime value. We discussed various features that may have an impact on the Lifetime Value.

   A. Total days in Lyft

B. The percentage that each driver works during prime time
C. Total distance driven within his career
D. Total time serving as a Lyft driver
E. Total number of rides
F. Onboarding date (categorized by months)
G. Average interval between request time and pickup time for every driver

IV. Preprocess

A. Apply backward elimination by removing features that do not have an impact on lifetime value
B. Standardize features by scaling to unit variance through the mean
C. Train/ Test split utilizing the ratio of 0.75 to 0.25

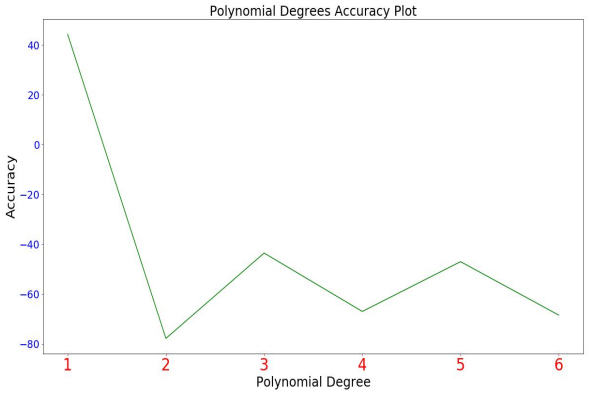
V. Implement different algorithms (fulfilled by supervised machine learning)

A. Random Forests

1. Random forests are an ensemble learning method for regression that operates by constructing a good number of decision trees at training time and outputting the class that is the mean prediction of the individual trees.
2. The algorithm yields a **96.550%** of accuracy. (Computed through MAPE)
3. Graph attached in the zip file due to the restriction of the pdf format. Through binary classification on feature's value, we can make predictions based on that.

B. Linear Regression

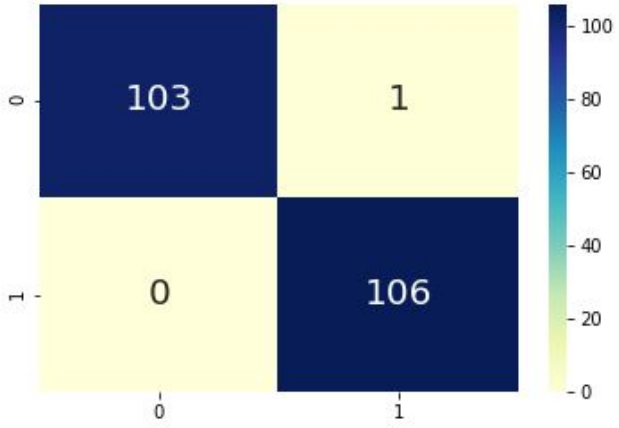
1. A linear approach to model the relationship between a scalar response and several explanatory variables.
2. The algorithm yields a 44.310% of accuracy(Computed through MAPE). Although the figure itself is not high enough, we can compare the value with higher degrees of polynomial regression. No matter which degree we choose, the accuracy is low enough that we can conclude the characteristics of the curve being more linear than polynomial. When polynomial regression



is implemented, although it satisfies the training data pretty well, the testing data are crying to death.

C. Perceptron

1. A classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector.
2. We categorize all Lifetime value into a negative one and positive one by whether they are more than average or not. And then we apply our features into the perceptron. And it gives us a 99.05% accuracy. Just as the graph shown below, only one false positive value exists. The result is satisfying!



3. Therefore, we can yield that with a higher lifetime spent in Lyft, more frequent prime time drives and longer distance covered, these factors can result in higher lifetime value.

Thank you