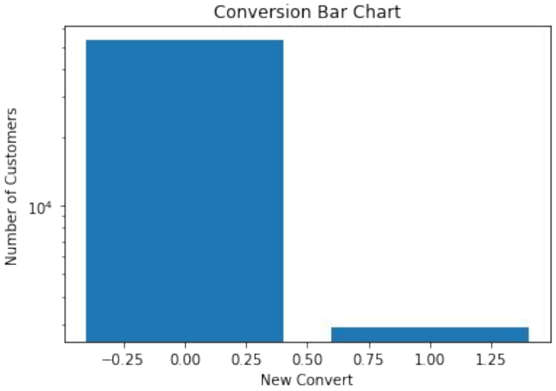
Wayfair: Using Data to Bring Customers Home

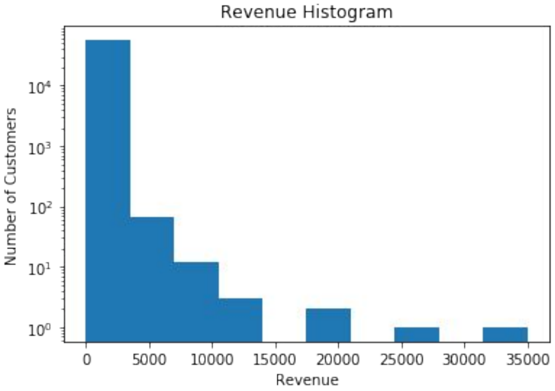
Travis R. Benedict

Exploratory Data Analysis

99.9% of columns and 91.7% of rows in the dataset contain at least one missing value. These values can not be ignored and were imputed based on the median value of their column

For both revenue and conversions the vast majority of entries are 0. This implies that the dataset is highly skewed.

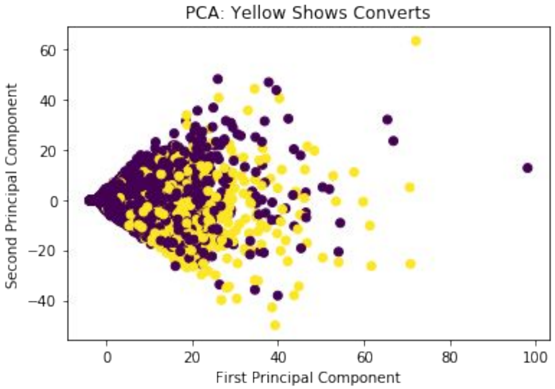


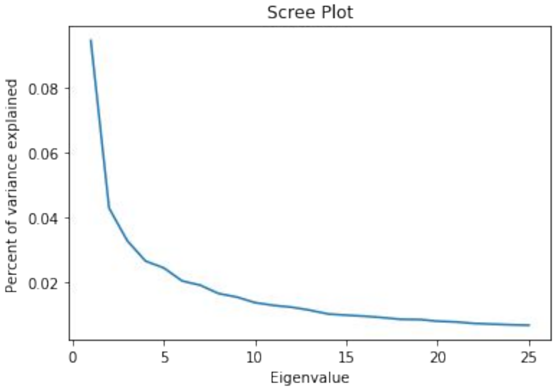


Exploratory Data Analysis

Principal component analysis offers little insight into this problem. The first two principal components account for less than 14% of the overall variance and do not offer a clear decision boundary for classification.

The first 25 principal components only account for 45% of the overall variance in the data





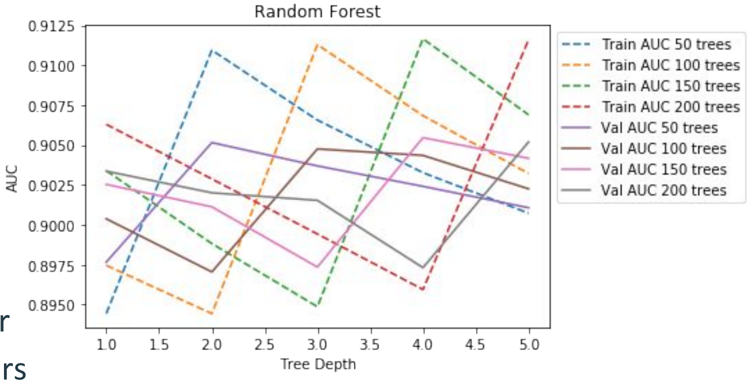
Data Overview

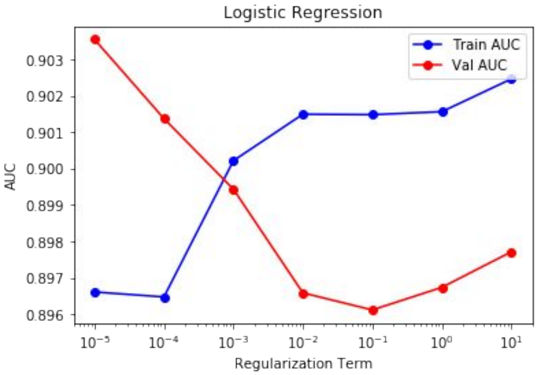
I split the labelled data into three sets based on a random sampling of rows: training validation and test.

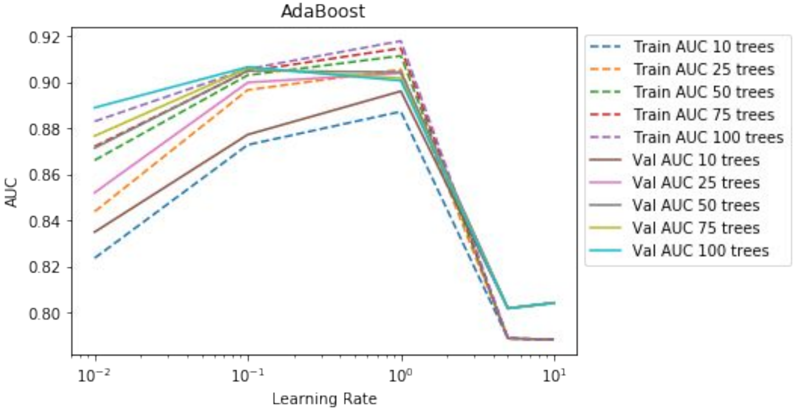
- 68% of the original data was used for training the models.
- 12% was used for validating hyperparameter choices such as the number of trees in a random forest.
- 20% was used for evaluating the overall performance of the models.

Classifier Selection

I performed a hyperparameter sweep on a validation set for logistic regression, random forest and an AdaBoost classifiers to try to determine the best model for predicting conversions.

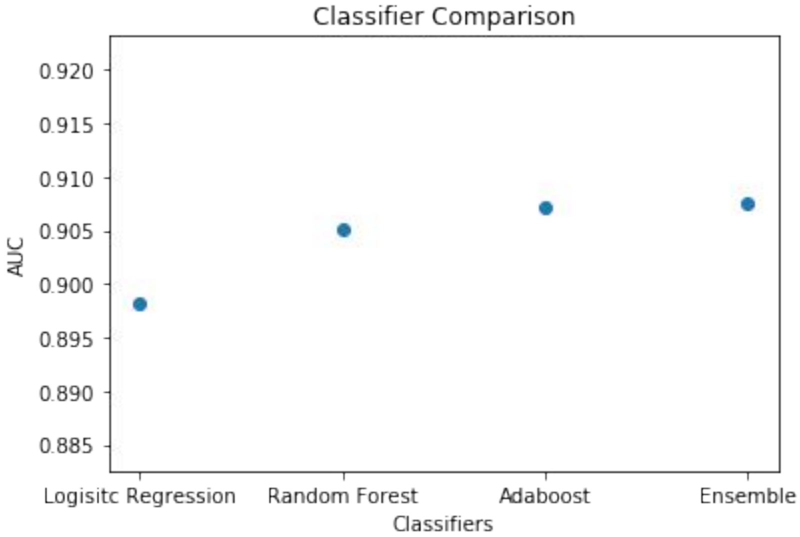






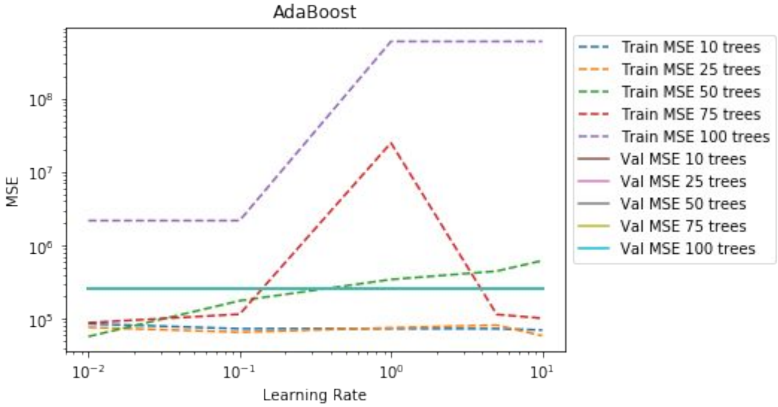
Classifier Selection

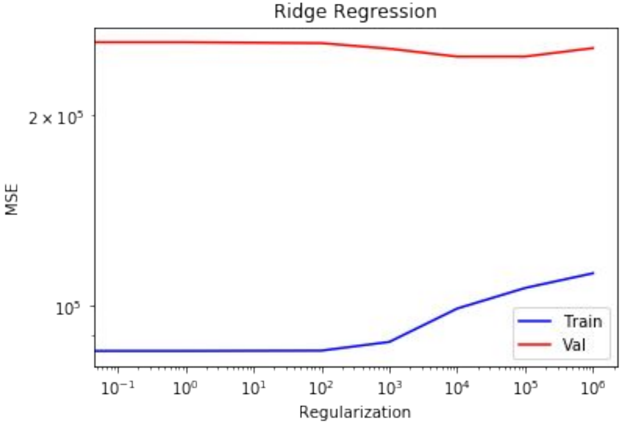
Each of the three classifiers was able to achieve an AUC over 0.89. Because of this I decided to ensemble their results together. In the ensemble each classifier is given an equal voting weight, meaning that if at least 2 out of 3 models predict a customer to be a convert they will be labeled as such. This method resulted in a slightly higher AUC on the test set than the AdaBoost classifier alone. I also expect the ensemble to generalize better since it will have a decreased variance.



Regressor Selection

To predict revenue I first added conv_30 as a predictor for the training data and its predicted value for the validation data. Then I performed a hyperparameter sweep for Ridge regression and AdaBoost regression models. Ridge regression was chosen over standard linear regression because its penalty term should help the model generalize better. I found that the best AdaBoost model had a slightly higher MSE than the best Ridge regression model, therefore I selected the Ridge regression model as the model for my submission.





Future Improvements

With a better computer and more time I would try to expand the features into a polynomial space. This may yield more accurate predictions, but would also need some sort of dimensionality reduction applied since most features would be nonsignificant.

I would apply a more clever strategy to imputing the missing data than simply taking the median of the column. Sklearn's Iterative Imputer seems promising but is only currently available in the developer release.

With more data or simply a more complete dataset I would try to apply a neural network to both the regression and classification problems. I did not attempt to use one for the current dataset because it would likely result in severe overfitting due to the relatively small size of the dataset. With so many missing values and a naive imputer augmenting the training data did not seem worthwhile.