Santander Value Prediction

A Kaggle Competition

The Goal

Santander is looking to identify the value of transactions of each customer they interact with. The purpose of this exercise is to develop a regression tool that Santander can use to estimate the value of an account or transaction then use that tool to personalize services for their customers.

Exploratory Data Analysis

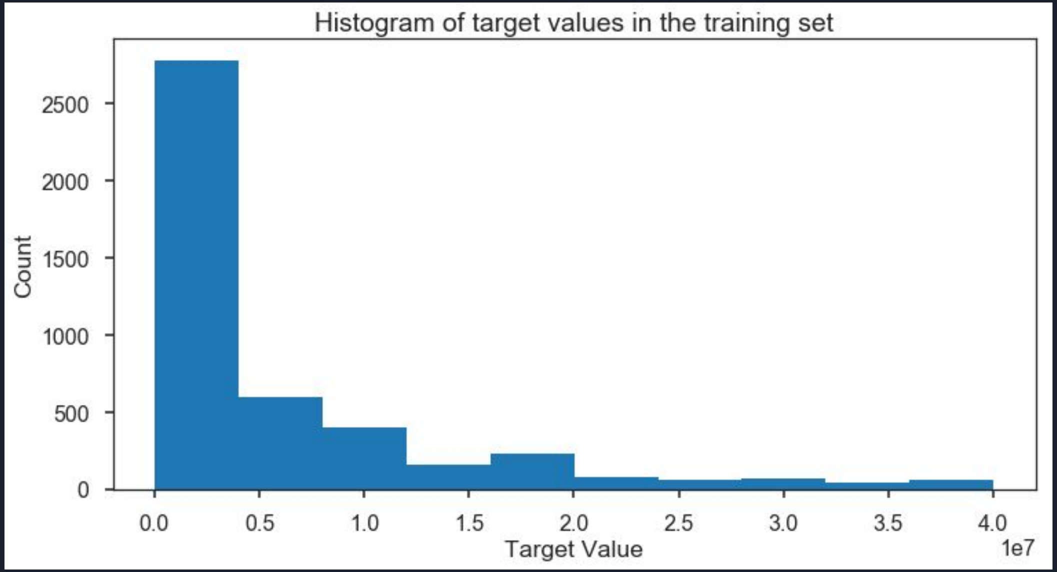
Data Cleaning

- First we ensured that there were no null values within both the test and train sets.
- Next we compared the column names between the test and train files to ensure consistency.
- Finally we checked the shapes of the dataframes to ensure they were the same as noted on kaggle.

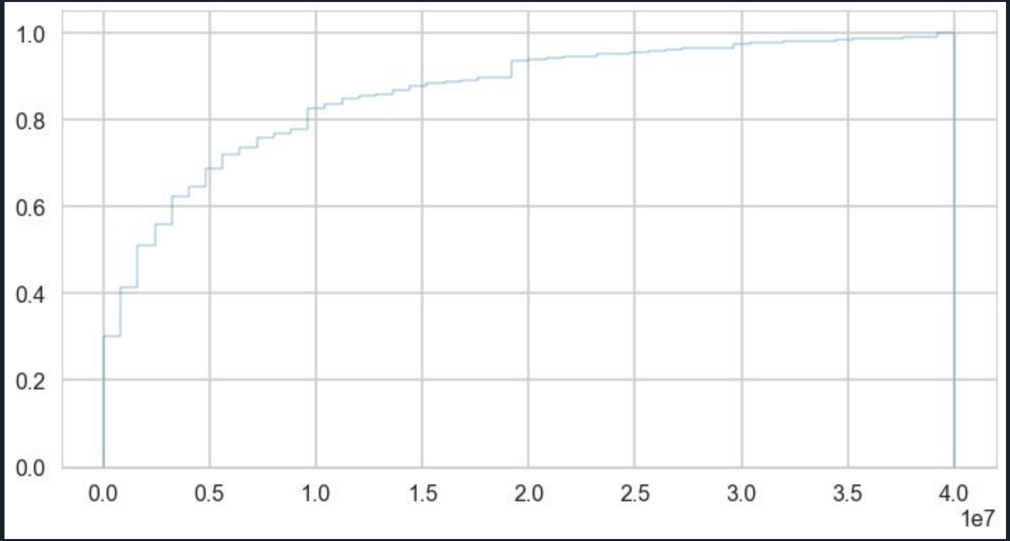
Histogram of target values

This plot seems strange at first But once you realized it is scaled by 1e7 it makes perfect sense.

Most customers have valuations fewer than millions of dollars, while very few have large accounts. Many of the large accounts could be tied to business partners, which would form up well with Santander's personalization goals.



Cumulative Plot of target values



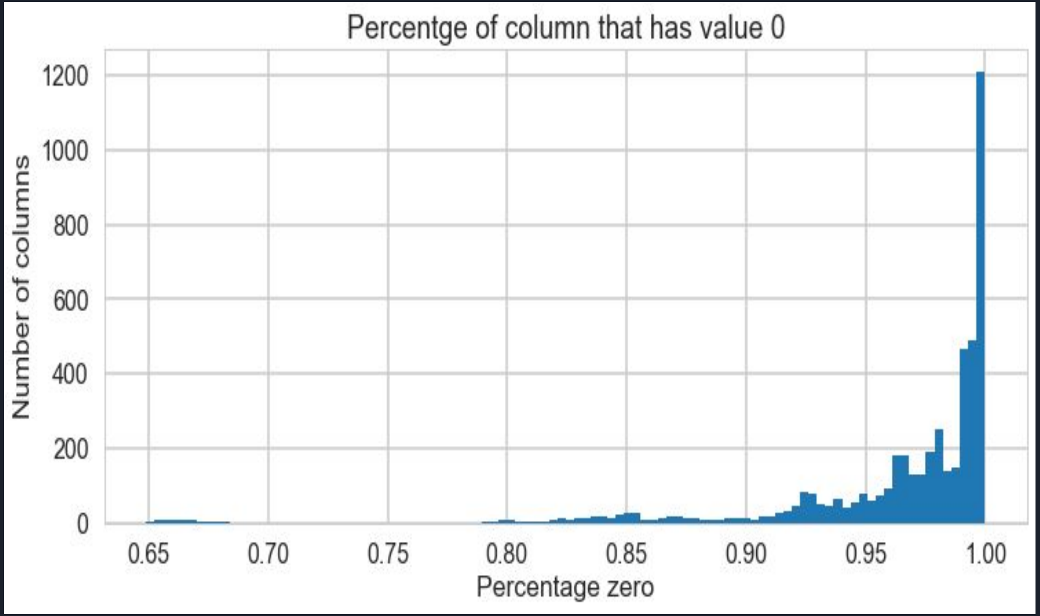
The cumulative plot illustrates how accounts with valuations under 1e7 make up 80% of the valuation of Santander's customer base.

This showcases the importance of individuals and small businesses to Santander.

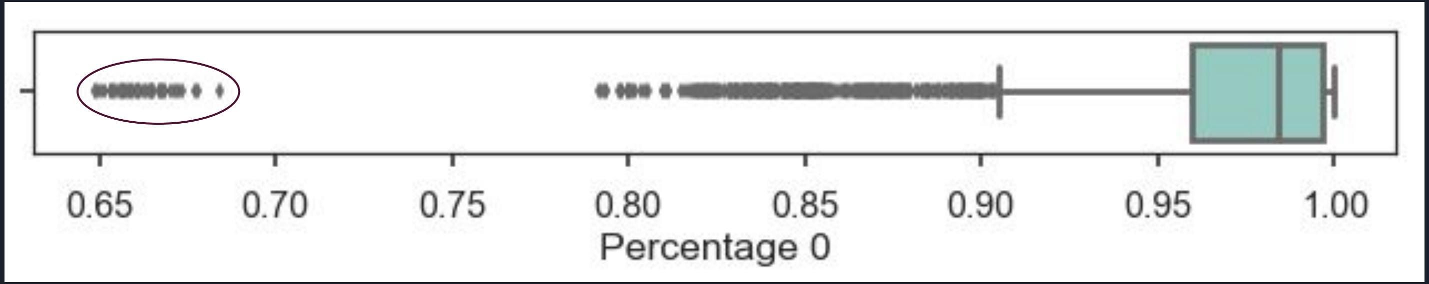
Extracting value from a sparse dataset

In the competition announcement Santander noted that the data was both anonymised and sparse. The first of which is to protect user data, the latter is a by product of the many avenues of transaction available to Santander's customers.

The plot to the right gives a first look into the situation at hand.



Box plot of percentage of zero value by column



The data is incredibly sparse, but there are clusters of more densely populated columns. The cluster of values with fewer than 70% zero values will be our focus. It contain 40 columns of data while providing the most useful information and reducing model error.

Machine Learning

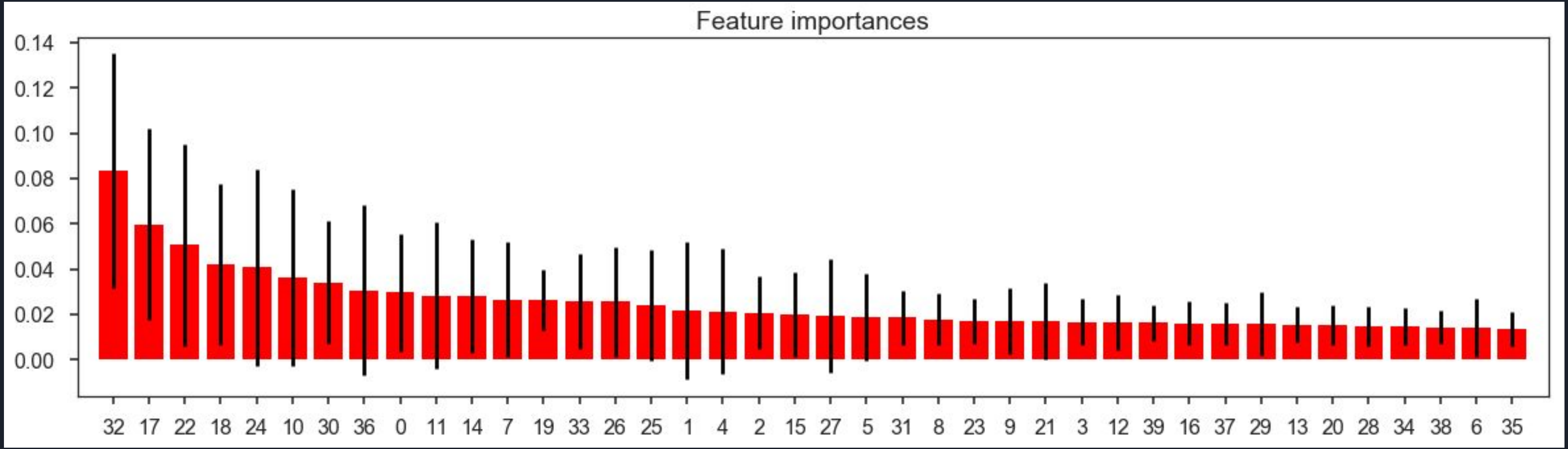
Comparison of Different Regression Models

Model	Error (RMSLE)
Linear Regression	1.68795
Support Vector Regressor	1.75909
Lasso Regression	1.68965
Default Random Forest	1.53282
Tuned Random Forest	1.44935

After Random Forest Regression proved to be the best base model, parameters were tuned using Randomized Search Cross-Validation.

Feature Importances

As the data was anonymised analyzing feature importances leads to fewer revelations in understanding of the problem. However it did prove useful. As you can see below a few of the features proved to be significantly more useful than the rest of the features. Unfortunately this was due to a data leak in the competition that made the target value appear in several columns. The feature importances plot helps give evidence to the claim which was later confirmed by Santander.



Next Steps

As of July 14th, 2018 there was a revelation within the kaggle community that there was a data leak where the target values appeared as values within certain columns of the data frame. As of July 16th, 2018 Santander has confirmed this and is currently working on a resolution.

This may involve removing the problem from the final scoring dataset, and as such tuning the model more to match highly important features may prove to be detrimental in that case. Thus I'm planning to submit the result from the tuned random forest regressor on the sub_seventy column set as my competition result.