Machine Learning Engineer Nanodegree

Capstone Project

Definition

Project Overview

For every business, customer relationship is most critical. Customer satisfaction is a key factor for success of any company. Even though company tries to collect customer satisfaction through surveys and polls. Most of the users don't respond to these. So, there is a need to come with new ways to implicitly measure customer satisfaction. Due to recent advancements in Big Data and machine learning, companies can store large amount of data about the customer interactions. This data can be analyzed using machine learning techniques to get an estimate of customer satisfaction. Santander bank is trying to identify unhappy customers using machine learning techniques. They posted this problem as a completion in Kaggle. They have provided both training and testing data which can be downloaded and analyzed.

Problem Statement

Main goal of this project is to predict how unsatisfied the customer with Santander bank is. Tasks of the projects are as follows:

1) Download the training and test data from Kaggle site.
2) Pre-process the data and make it more ready for the model.
3) Build a model and fit it on the training data.
4) Using the trained model get the predictions of the test data.
5) Submit these predictions and get score of the model.
6) Based on the results tune model to perform better.

Final python script should be able to take new customer data and give a prediction of how unsatisfied the customer is.

Metrics

In any business most of customers will be happy with the company. So, percentage of unsatisfied customers will be pretty less. Accuracy will not be a correct measure for these kinds of problems. Area under Receiver Operating Characteristic (AUROC) curves will be more appropriate measure in this situation. ROC curves are constructed by taking ratio of Sensitivity (Identification of truly unsatisfied customers as unsatisfied) to Specificity (Identification of Satisfied customers as satisfied) at different threshold levels. A ratio of one means all the unsatisfied and satisfied customers are identified correctly.

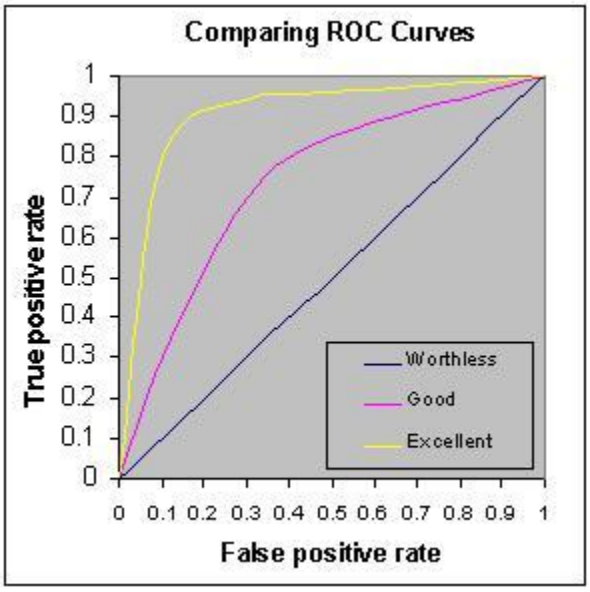

Figure 1: ROC CURVES

From the above figure we can see that, Yellow curve which is near top left performs the best and Red curve performs average. So, an Ideal model will be on the top left most, covering an area of 1. So, AUROC value of 1 is highest score that can be achieved by the model and value of 0 indicates worst model. 0.5 is the value of the benchmark model.

Analysis

Data Exploration

- Both training and testing data sets are provided separately. Training set has 76020 rows and 371 features including Target variable. Test set has 75818 rows and 370 features. Training data has to be separated into training features and training labels for supervised learning. There is also ID variable which is just a numerical representation of the customer that has to be extracted from both training and testing data.
- There are no missing values in both training and testing data sets.
- 4807 Duplicates were found in training data after extracting the customer Id.
- Maximum value in train data is 9999999999 and minimum value is -999999.These values deviated from other data by two larger amounts. So, there two values are considered as outliers.
- All the features of the training and testing set are numerical. Of these 258 features are discrete (‘int64’) and 111 are continuous (‘float64’) variables.

- Difference between common stats like mean, STD, min, max is very large. If we train the data with this, model will give high weightage to features containing large values.
- Number of features is very large for such a small dataset. So, Feature Selection has to be done for removing unwanted features.

Exploratory Visualization

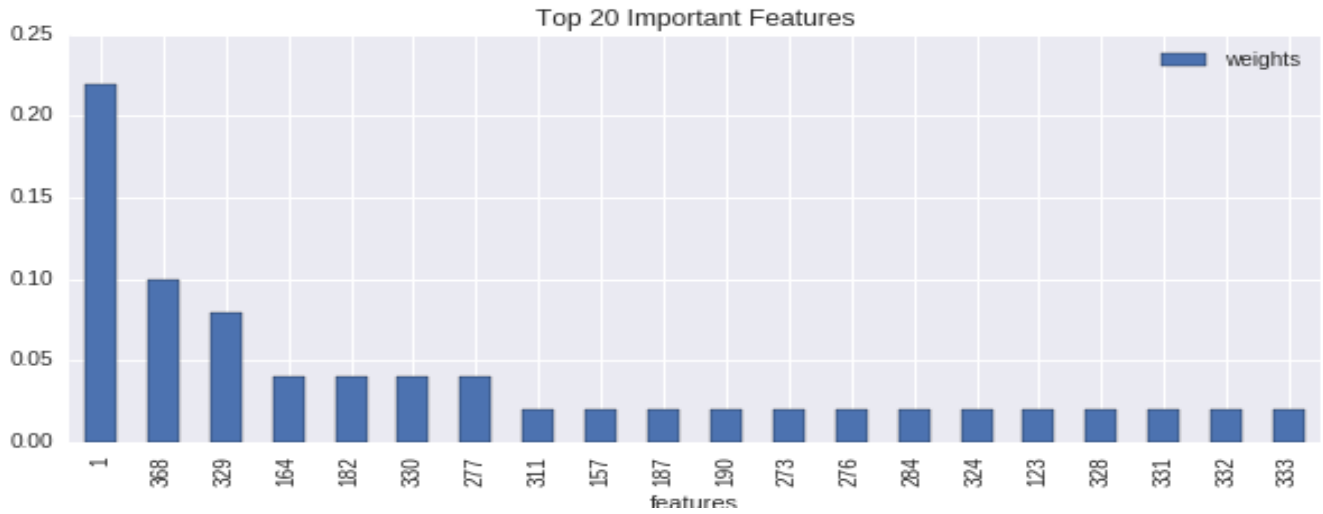

Figure 2: Top 20 features weights

These are the weights of the top 20 features that contribute the most in predicting the output label. We can see that column 1 has the highest weightage of nearly 23 in deciding the output and after 7 features nearly all the rest features provide a weightage of about 2%. This indicates that while predicting we should take into account more features so that we don't miss the affect caused by these features.

Above picture represent the box plot of top 20 features of the normalized trained Data.

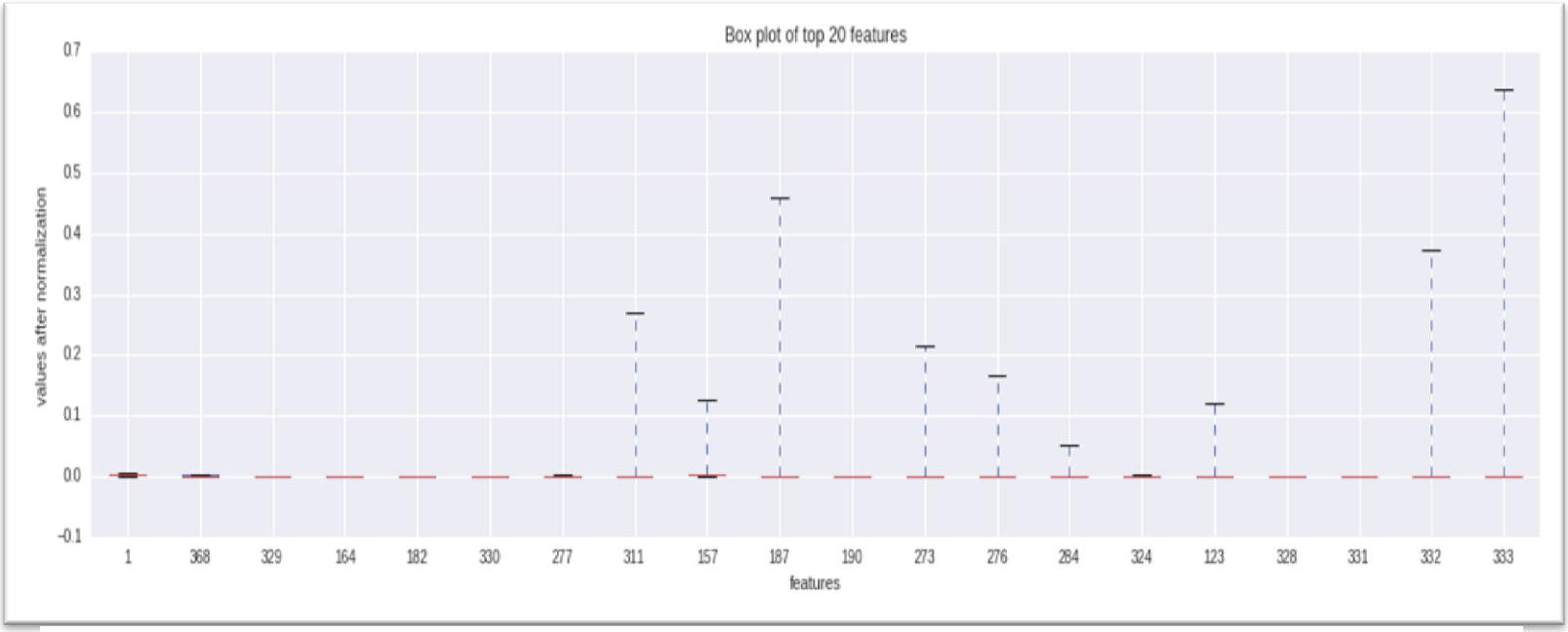

Figure 3 : Box plot of top 20 features

As we can observe some of the features like 311,187,333 have values which are far higher than the rest of the data. These we can say as outliers. But as we really don't know what these features represent we cannot remove them without complete understanding (Feature description is not provided along with data set.).We cannot assume these as noise or miscalculated data, because it is common in bank system for very few individuals to have account properties which are far different from others. Above boxplot is perfect reflection of that. That two first 7 features contribute most to the final predictions, so few different points in these features are not going to affect output prediction by large factor.

Algorithms and Techniques

Random Forest

Random forest or random decision trees are an ensemble learning method which is used in classification and regression problems. By selecting random features at each node random trees are generated. Using untrained records (random sample with training set), accuracy of each tree is measured and appropriate weight is given to each tree. Top n trees with highest weight are selected and remaining trees are discarded. Final prediction of the model is evaluated by taking mean of predictions of each tree multiplied by their weights.

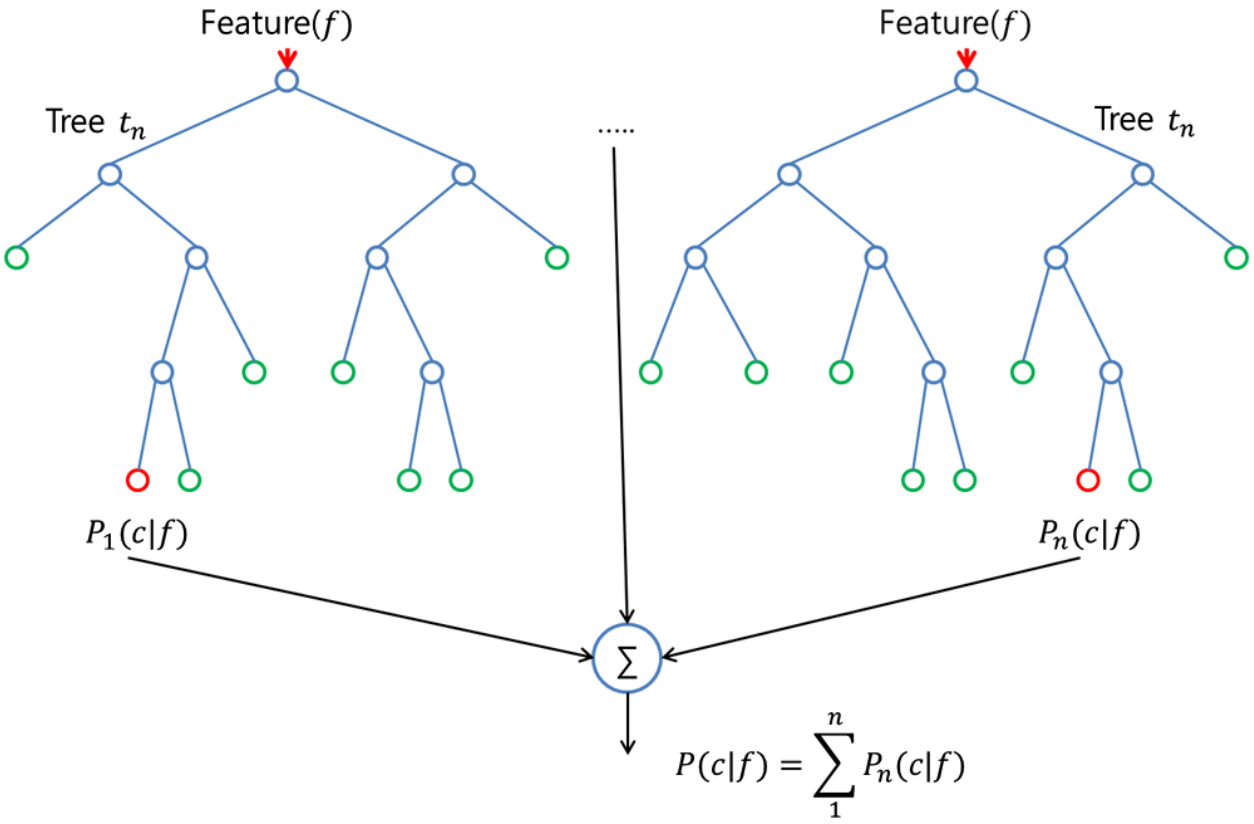

Figure 4 : Random Forest

Random forest are robust to outliers, they can handle abnormal data pretty well. After extraction of 95 percentile important factors, we have just 26 variables in the data base with 71213 records. Random forest performs exceptionally well when number of variables is far less than the records.

As target variable here is dichotomous (binary), it would be compelling to use logistic regression. But as show in the figure 3, out data contains some abnormal (outlier) data. Logistic regression will provide more importance (weightage) to these features which will result in biased results. So, logistic regression is not suitable for training this dataset.

GridSearchCV

Grid Search is a model optimization technique where different models are generated by varying the parameters of the selected method or model. GridSearch exhaustively generates all possible combinations of parameter values specified in the param_grid. For example, in SVM, if we provide param_grid = [{'C': [1, 10], 'gamma': [0.001, 0.0001], 'kernel': ['rbf']}].It will check model with 4 different combinations of C, gamma values with 'rbf' kernel. Of all these combinations, we are fine tuning the parameters (considering all possible combinations) of the selected model (learning algorithm) to get the optimum

results. Model with highest performance (selected score measure) is considered as the final model.

Benchmark Model

In any business entity, most of the customers will be happy with the service (except for few). Same is the case with the Santander bank, most of the customers of the bank are satisfied. A quick analysis of the train data shows that 96.04 are satisfied. So, will have any accuracy of 96.04 if we predict all the customers are happy. This will be a good benchmark solution for this model. Submission assuming every customer is satisfied gave AOC (benchmark) score of 0.5.

Figure 5 : Benchmark Model AUROC Score

Methodology

Data Preprocessing

- Presence of missing values in both training and test set are checked. There are no missing values.
- CustomerID in training data is not required, so training CustomerID is dropped.
- Presence of duplicate records in the training data is checked and 4807 duplicates are removed.
- Maximum and minimum values of the training data are checked to look for any abnormalities.
- Records containing value of 9999999999 and -999999 are removed from training data. These values are noisy data.
- Now labels (target variable) from the train are separated and stored separately.
- CustomerID is separated from test data and stored separately.
- Great difference in value distribution between different features is observed in training data. For example, most of the features has a mean of 0 but minimum value in Var38 is 116975.3
- So, data is normalized along each column to achieve zero mean and equal standard variance.
- Now, every column in training and testing data has near zero mean and small standard deviation. This prevents the model from giving high bias to certain variables.
- We have 369 features in training and testing data and only around 70000 records in both.
- So, it will be difficult to train model with such large features on small data. So, feature selection is performed on both training and testing data.

- A model is chosen with adabooster classifier. Model is trained with train features and train labels.
- Weightage of each future in predicting the outcome is extracted from adaboster and stored in pandas Dataframe with corresponding features. DataFrame is sorted based on the feature weights in descending order.
- By summing up cumulative weights, features which add up to 95 percent weightage were extracted. There are total of 26 features.
- Similarly important 26 features are extracted from testing data.
- Now both training and testing data is processed and prepared for training a model.

Implementation

- Target variable in the training data is binary. So, two well-known algorithms that can be used for this case are logistic regression and Random Forest.
- From the box plot in figure 3 we can see range of values in few features is pretty high. Few values are really higher when compared to other values. In these cases logistic regression will give high bias to these features and logistic regression is not good at handling noisy data.
- Lack of target variables for testing data made is difficult to clean and preprocess the data difficult. There might be presence of duplicates and other anomalies in the test data.
- Under these conditions it would be appropriate to consider Random Forest as it can handle noise very well. Having only 26 features and nearly 70000 records will make the training easy.
- Default RandomForest with random state 50 is chosen to train the model at first, with number of estimator in RandomForest is 50 and Max_depth as None.
- After training ,test set features are provided to get corresponding probability predictions
- These probabilities are used in construction ROC curves from which we calculated AUROC value.
- This model gave an AUROC score of 0.617626 which is not enough.

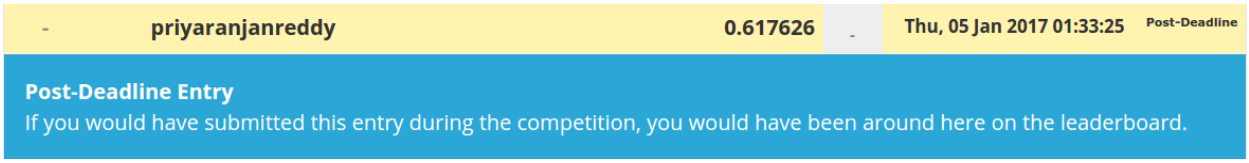

Figure 6 : Default RandomForest AUROC Score

Refinement

- Even though result from the default random forest is higher than the baseline model, it is not enough. This model can be improved by tweaking n_estimators and Max_depth parameters.
- For this purpose we use GridSearchCV , which generates all combinations exhaustively between given parameter values.

- Parameter chosen for GridSearch are n_estimators : [1000,1200] , Max_depth :[13,20] and 10 K-fold cross validations are chosen for each parameter and 'ROC' as the scoring parameter. Parallel Processing with 4 jobs is initiated to decrease over all runtime.
- Best parameters for refined model are n_estimators = 1200 and Max_Depth = 13

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=13, max_features='auto', max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=1200, n_jobs=1, oob_score=False, random_state=50,
            verbose=0, warm_start=False)
```

Figure 7 : Best Estimator

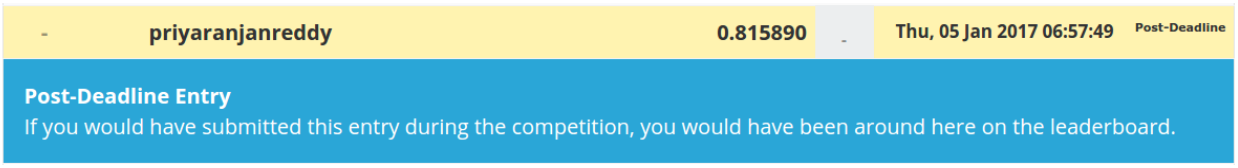

Figure 8 : AUROC score of best model

Results

Model Evaluation and Validation

- Classifier.best_estimator gives the best combination of the GridSearch parameters which gives highest AUROC score.
- We can see that max_depth of the best classifier is 13, so the depth of the tree is controlled to avoid overfitting.
- With 600 estimators model was complex enough to capture the patterns in noisy training data. More number of estimators would have made model more complex there by resulting in over fitting.
- Choosing the best parameters a random forest classifier has been initiated.
- Noise is created in training data by adding probability of standard deviation (-1 to 1 range) to 10% percent of values for each corresponding feature.
- Model was robust enough to tolerate little noise in the training data and gave same results.
- Comparing the result before sensitivity analysis and after, it can be said that model is robust enough and can be trusted.

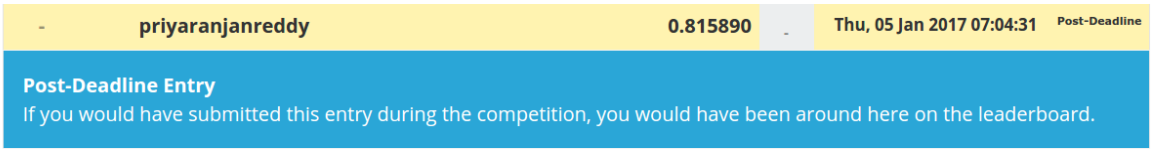

Figure 9 : AUROC score after Sentivity

Justification

- Benchmark model produced a result of 0.50 where are robust and refined final model gave result of 0.8158
- This is a 31% increase in the AUROC score, the top model in the competition AUROC score is about 0.8453.Considering this, it can be stated that Final model has performed well and even after sensitivity analysis model performed well.
- Final model can is significant enough to identify unsatisfied customers.

Conclusion

Free-Form Visualization

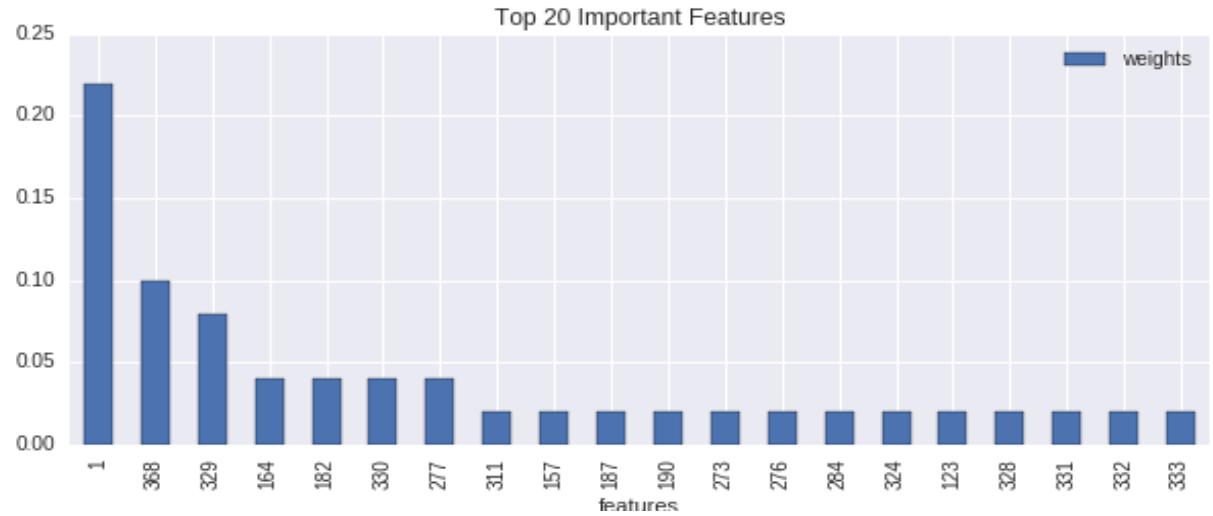

Figure 10 : Top 20 Important Features

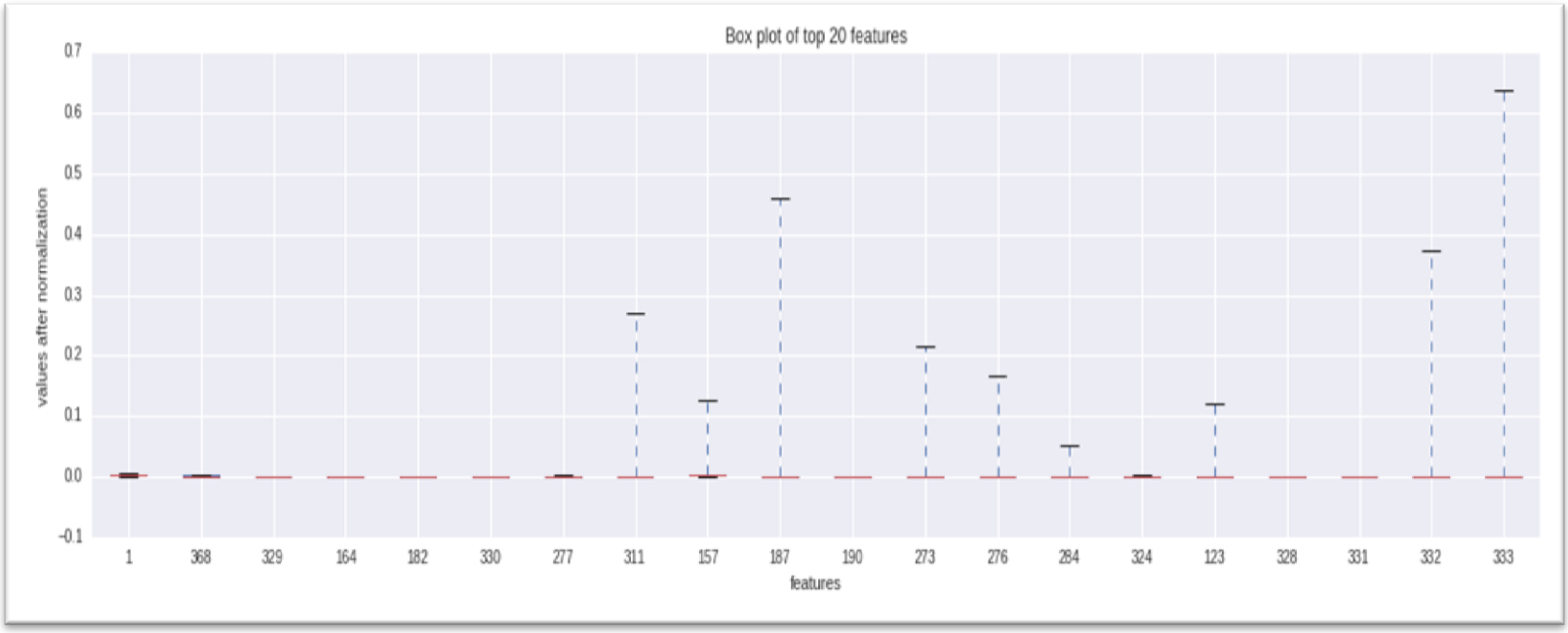

Figure 11: Box plot of top 20 important features

- Figure 8 plots the Weightage given to top 20 features and figure 10 represent the box plot of top 20 features.

- In the weightage distribution we can see that, weights after 7 features are around 2%. If not for this distribution it would be easy to overlook the last 20 features and just take first 7 features as final features. Last 15 features contribute about 30% of the weightage. This was one of the reasons to take 95% of the features weights for final model evaluation.
- In figure 11 we can see in the box plot that , few features after first 7 features contains values that are far higher than other. These values are outliers and it is normal to have these kinds of outliers in banking data. But, when figure 10 and figure 11 are compared, it can be seen that features containing outliers have less weightage. Because of this we can expect the disturbance caused by outliers are minimal
- Considering presence of few noisy data and outliers it was evident that RandomForrest was a good choice as supervised machine learning algorithm.

Reflection

Complete process of the project can be summarized as follows:

1. Collecting training and testing data.
2. Checking for missing values in the data.
3. Removing unnecessary features
4. Checking and removal of any duplicates.
5. Extracting features and target labels from training and testing data.
6. Getting common statistics like min, mean, median, max, STD to get some insight in the data.
7. Applying normalization as data contains improper distribution of features values.
8. Getting the weightages of each future.
9. Plotting feature weights and box plots for top 20 important features.
10. Deciding Random Forest would be a better supervised learning algorithm from the plots.
11. Extracting top 95 percentile weightage features.
12. Establishing a benchmark model and getting benchmark score.
13. Training a default Random Forest Model and getting AUROC score of the model.
14. Using GridSearchCV to tweak parameters.
15. Finding the best parameter from GridSearchCV parameter combinations.
16. Doing sensitivity analysis to find if the model is robust enough.
17. Confirming model is robust enough by comparing AUROC scores before and after sensitivity analysis.
18. Finalizing the model.

Interesting aspects:

- Feature scaling enable to select just 26 columns providing 95% weightage than considering whole 369 features.
- It was interesting to note that less than 10% of the features were able to capture the whole essence of whole data.

Difficult aspects:

- Even though GridSearchCV helped in providing the best parameters, It is an exhaustive task to search through the parameters by just trial and error.
- Missing target variables from the testing set prohibited more thorough processing of data and tuning of model.

With 0.815 AUROC score final model has ability to predict unsatisfied customers and procedure stated above can be applied to similar kind of problems.

Improvement

- By using ensemble of different algorithms like logistic regression, Random Forest, XGboost and bagging (providing weightage to each algorithm and combining them for the final result) will produce a better result.
- Analysis of the procedures followed by top 100 solutions indicates the usage of XGboost with combination of other methods. It shows that most of the top competitors use XGboost for most of their solutions which make is compelling algorithm to understand and implement.
- Highest AIUROC achieved in this completion is 0.8158 which is little better than the final model presented here.