BACKORDER PREDICTION

AMAR CHHEDA

ROHITH NAGABHYRAVA

SANKET KULKARNI

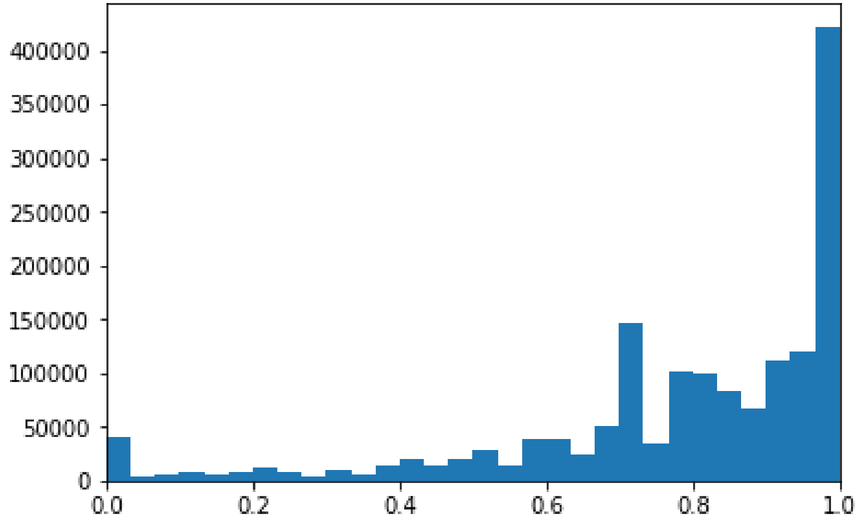
PRANAV GHODKE

INTRODUCTION

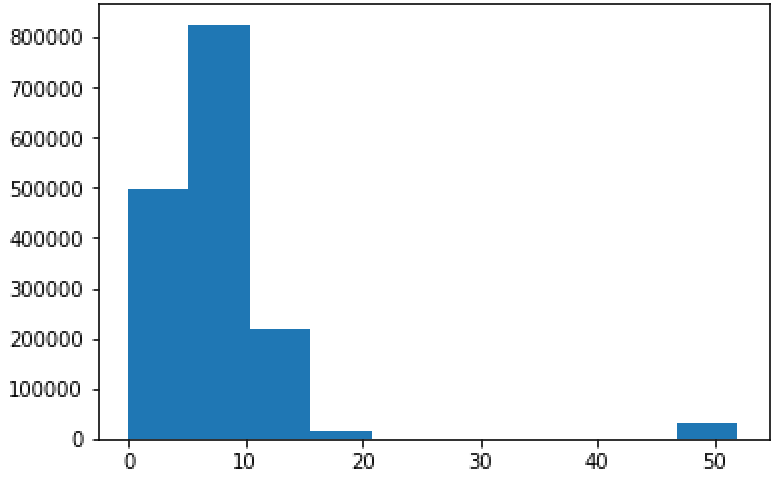
- What is backorder
- Is it good or bad?
- How to deal with backorders?
- Why backorder predicting is necessary?

Data Description

- Our data is 16,00,000 * 23
- It consists Numerical and categorical variables
- Data is majorly imbalanced
- The numerical features have different scales
- Features such as Lead time, performances have outliers and missing values ranging from 5-10% of total data
- The target variable is went_on_backorder which is categorical variable with Yes/No response
- Data is highly imbalanced - Yes response forwent_on_backorder is only 0.7%



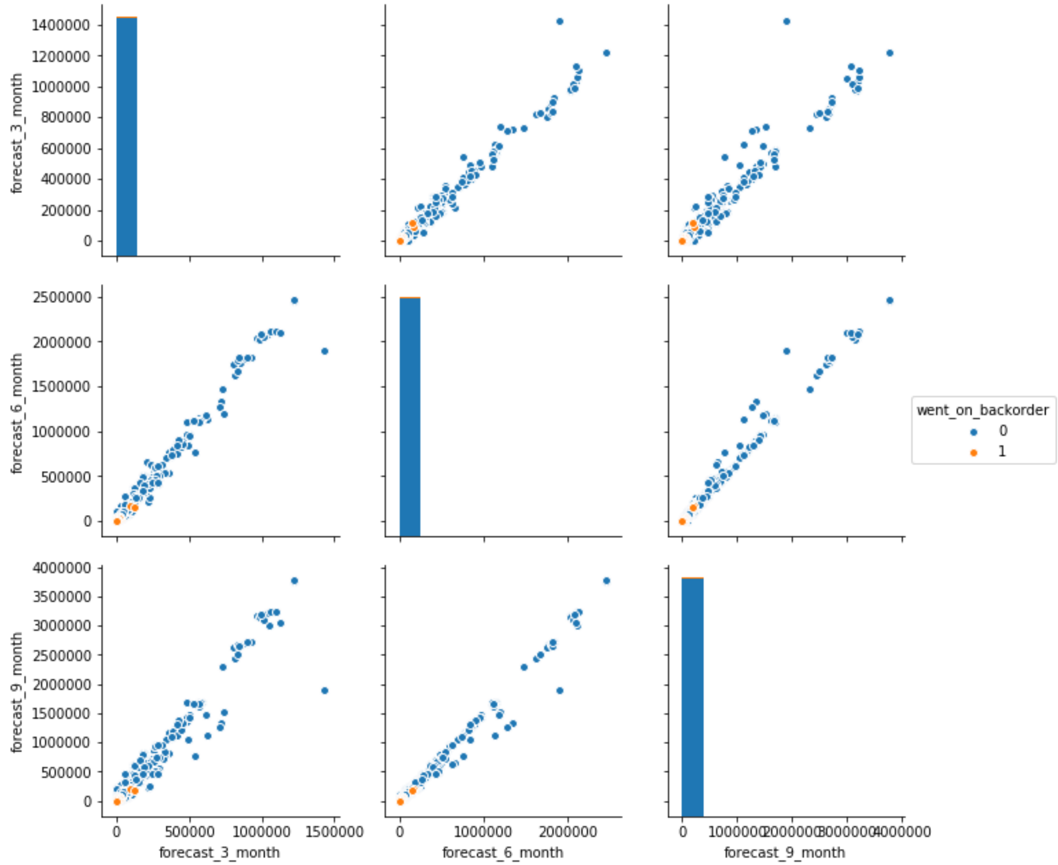
Distribution of perf_12_month_avg



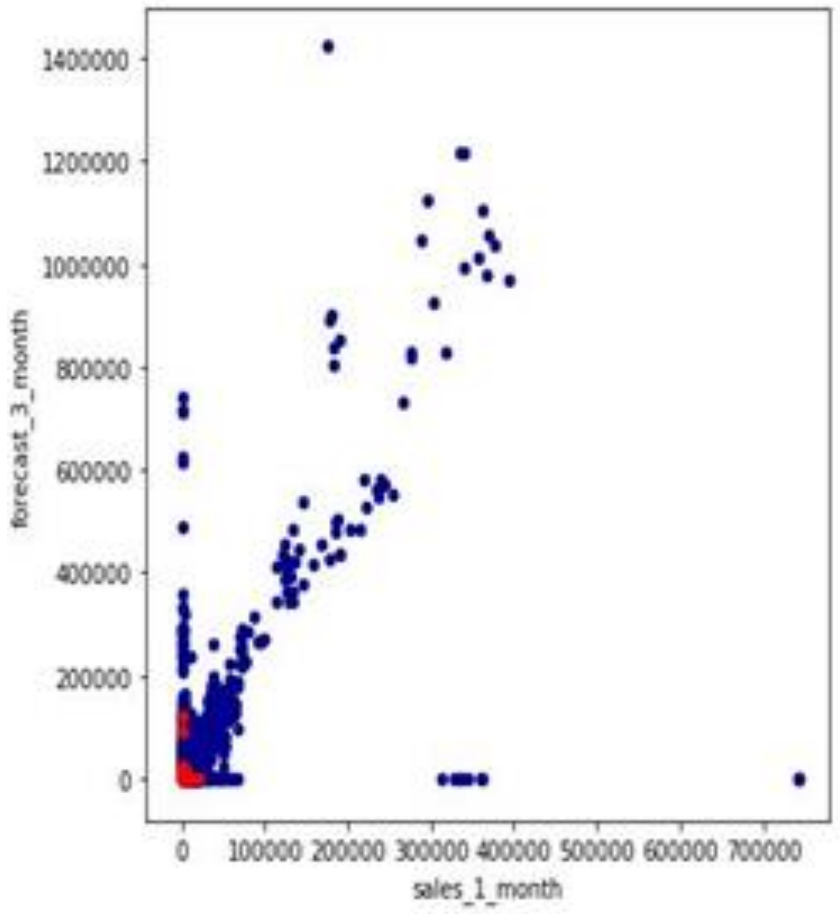
Distribution of lead_time

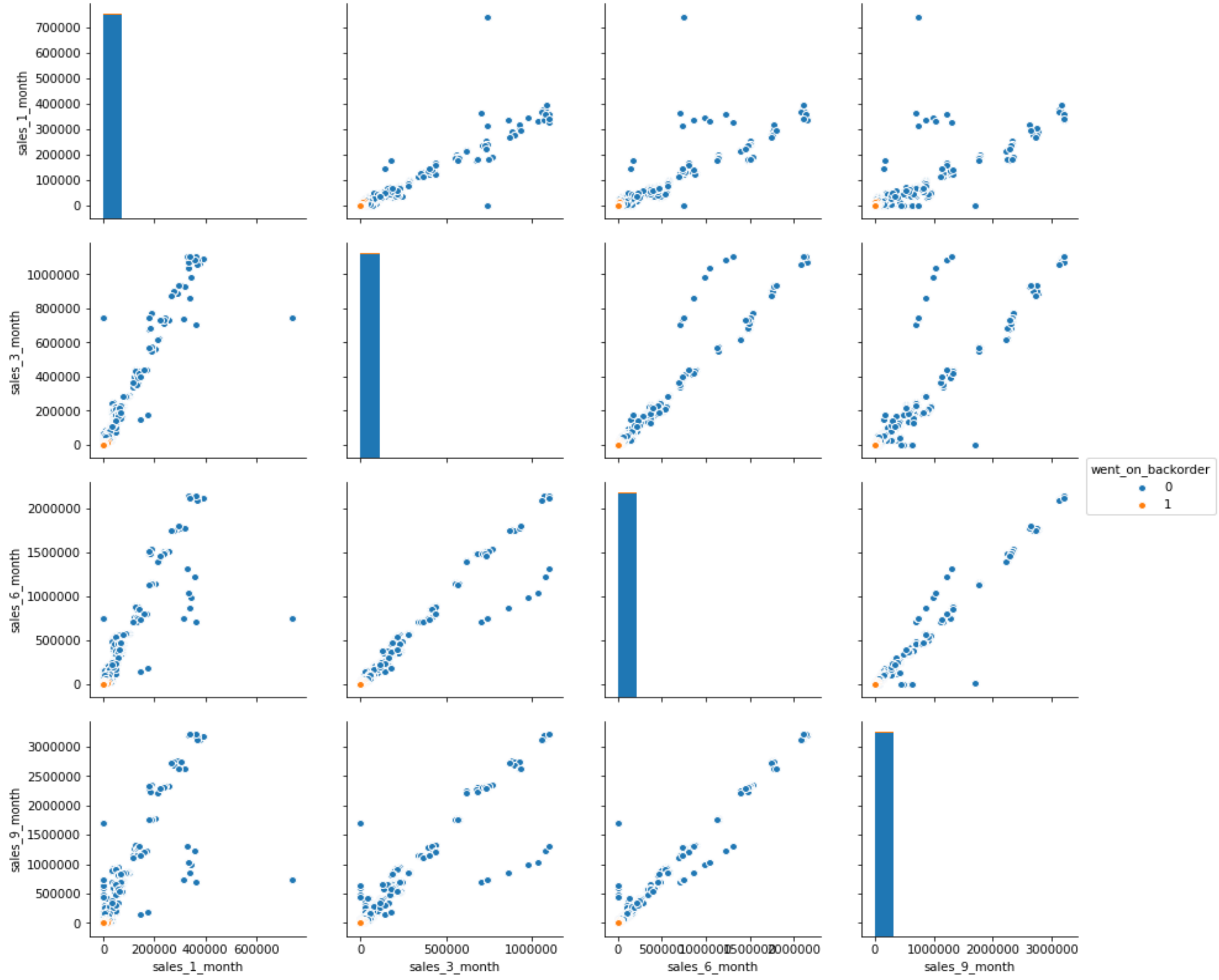
Feature Selection

- Features were selected by applying domain knowledge of Supply chain
- Forecasts and Sales were analyzed for 3, 6 and 9 months data
- We can see that the relationship between the variables are linear
- We also observed that the backorders happen only when the value of sales and forecast is very low
- Linear relation was observed between sales and forecast data
- Due to the good correlations and sufficiently linear relationships between these features we concluded that sales_1_month can represent all forecast and sales data



Feature Selection - Contd





Approach to handle Imbalanced Data (only 0.7% went on backorder)

- Oversampling
    - Minority class is randomly replicated.
    - Increases the overall size of the data
- Under Sampling
    - Randomly eliminating the majority class
    - This method help improve run time and the storage problems by reducing the sample size
    - There can be loss of potentially useful information which could be important
- SMOTE (Synthetic Minority Over-sampling Technique)
    - Used to avoid overfitting which can occur when replicating the minority class
    - SMOTE is found not effective for high dimensional data.

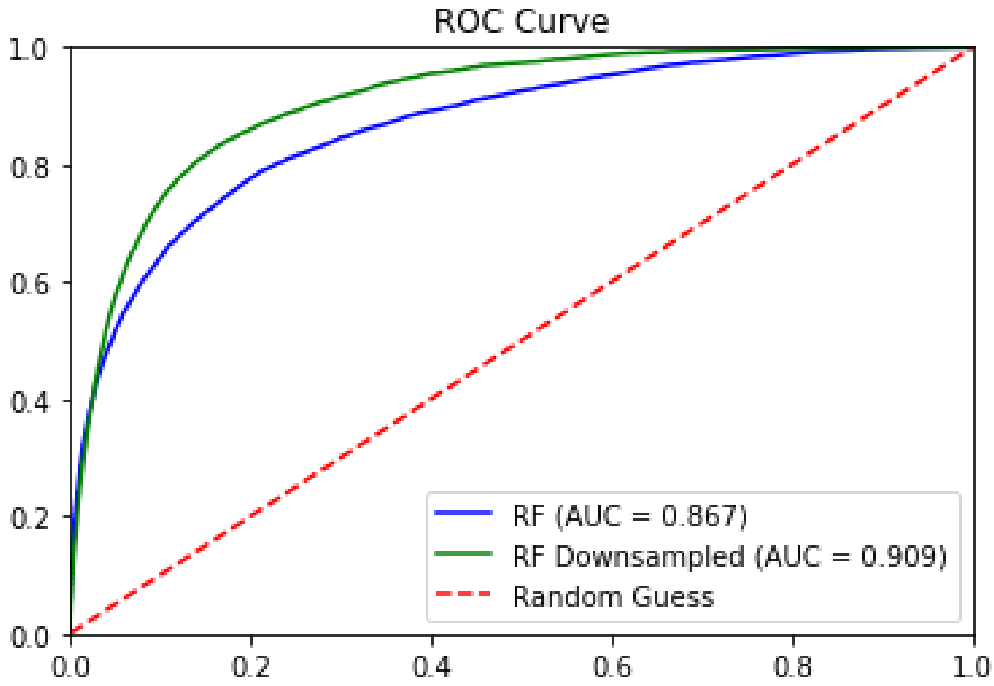
Algorithm Selection - Random Forest

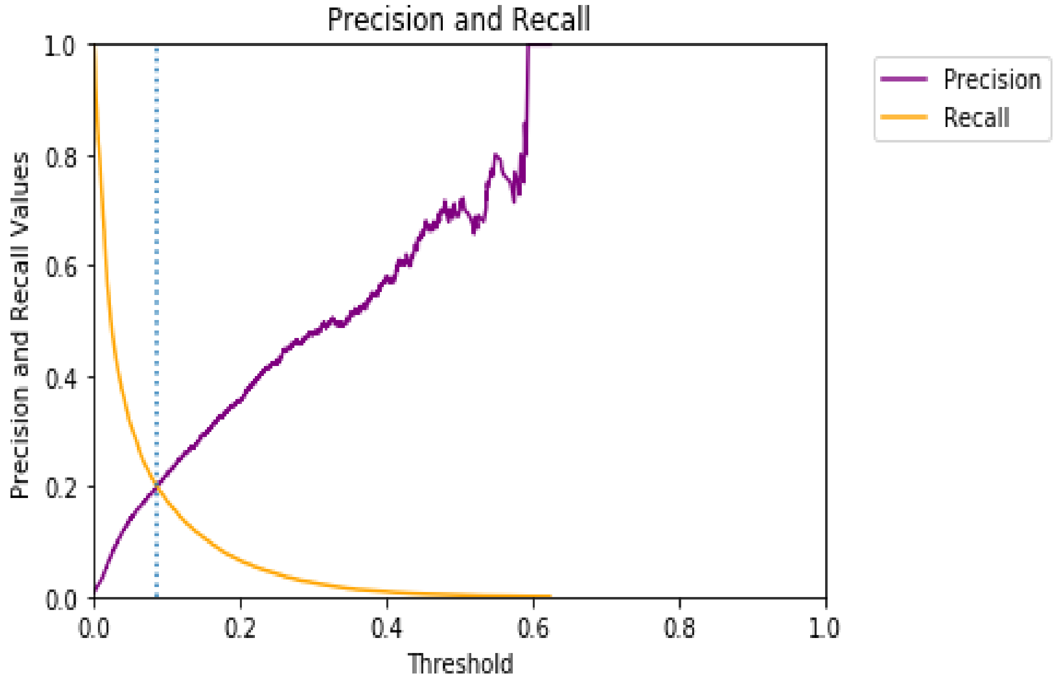
- Robust to outliers and missing values
- Perform well with large dimensional datasets
- Can handle thousands of input variables without variable deletion.
- Gives estimates of what variables are important in the classification
- We compared the model by varying number of leaves and the minimum support.

K-fold Cross Validation

- The original data is randomly partitioned into k equal sized subsamples.
- A single subsample is used as the validation data for testing the model, and the remaining k-1 subsamples are used as the training data.
- The advantage of K-Fold Cross validation is that all the observations in the dataset are eventually used for both training and testing
- Reduces bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set

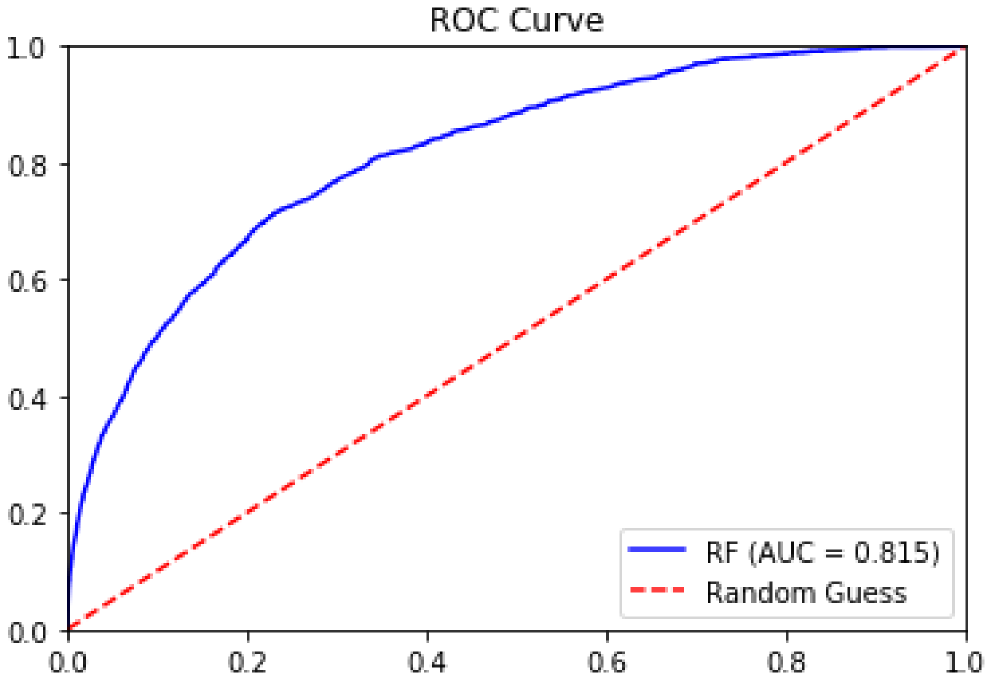
Downsampling Cross Validation Results

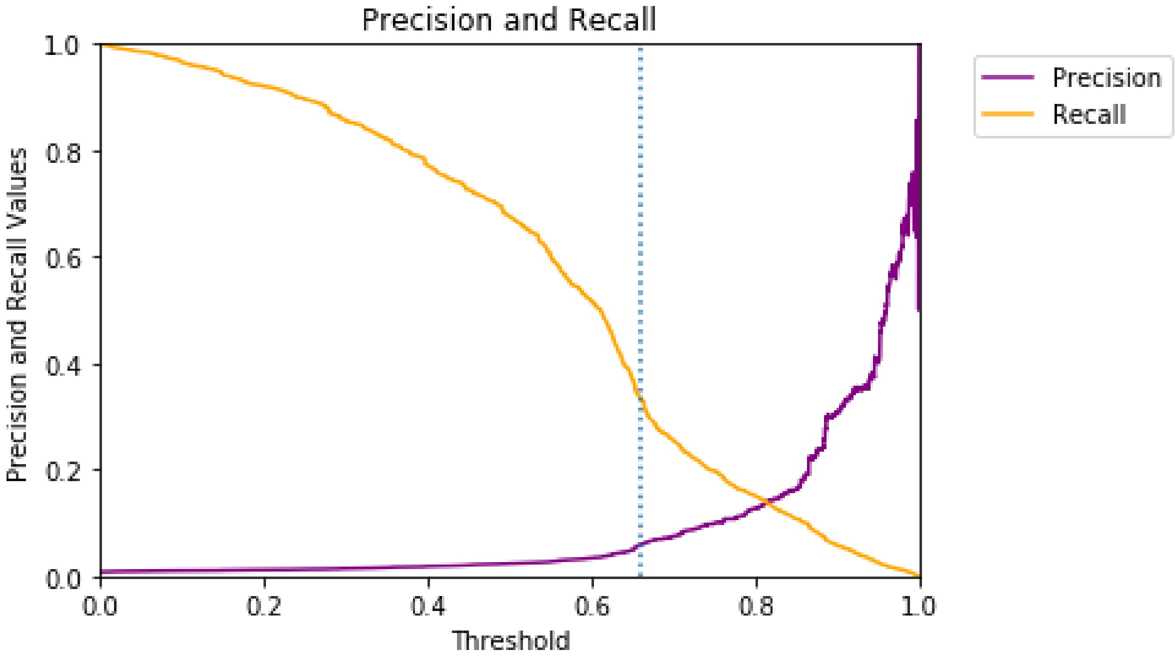




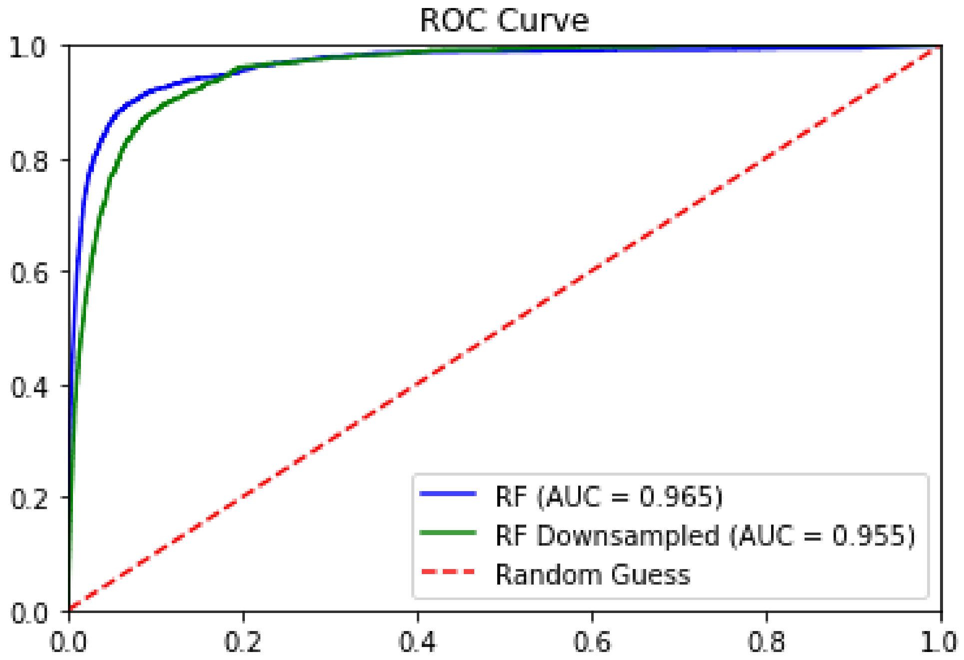
Number of estimators: 50 | Maximum features: 7 | Minimum leafs: 5

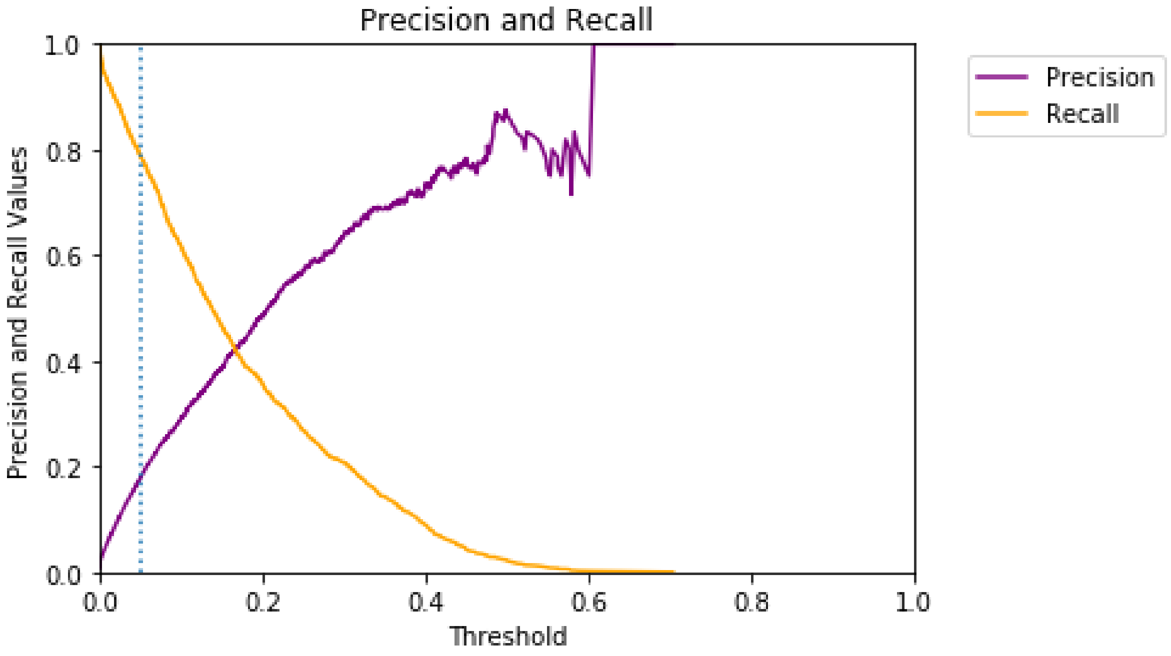
Results Using SMOTE





Results With All 22 Features





Interpreting The Results

- Why overall accuracy is a bad predictor?
- Why just ROC curve is not enough?
- Why Precision - Recall graph is important?
- Precision VS Recall
- Computational Time Complexity VS Results

Future Work

- K-fold Cross Validation before sampling
- Higher K values for cross validation
- Different sampling techniques.

Questions?

Thank You! :)