Applying Machine Learning Techniques to Mortgage Loan Default Prediction

11.26.2018

Mike Labadie

Overview

The Washington D.C. area is home to two of the most important organizations in home mortgage lending, Fannie Mae and Freddie Mac. There is also a large private sector services industry built up to these two large entities. As a result, there is no shortage of analysis of home mortgage loan defaults being conducted in this metropolitan area. This presents a ripe domain for employment opportunities as data scientist in Washington D.C.! My preliminary research has revealed that logistic regression is the most commonly used tool to predict mortgage defaults. With this exercise, I sought to identify other machine learning techniques, if any, that produce high quality mortgage loan default prediction models.

The goal of my exercise was to find techniques that will identify individual loans where the borrower will experience difficulty making repayment (i.e. default). I'm taking the perspective of a lender originally initiating a loan. Relative to many machine learning classification exercises, predicting mortgage loan defaults at the individual sample level (versus analyzing default rates at macro level) presents a difficult challenge. Less than 1% of mortgage loans will experience a default in normal economic situations. Further, human behavior and economic situations can be difficult to predict. Well qualified borrowers could face illness/death/job loss that could prevent them on making mortgage payments. Borrowers with lower initial credit worthiness could put previous financial instability behind them. Further, capturing information on a broad array of economic factors that affect a large number of borrowers is daunting.

With a brief literature review, I saw a wide variety of techniques being used to predict trouble with mortgage loans. Infamously, large mortgage focused institutions associated with the federal government produced flawed models, which relied heavily on regression techniques, heading into the financial crisis of the late 2000s. I've also seen more recent research that used various machine learning techniques with quality results. Much of this modeling centered around predicting defaults during the life of a loan within a smaller window, such as one to three years. These models often highlighted that the age of the loan is one of the strongest predictors of default. These models also relied on timely econometrics such as unemployment rate. While I was impressed by the techniques and results, I focused my efforts on modeling loan defaults from the time the loan is originated.

Data

Fannie Mae is one of the most important institutions in the mortgage lending system. It purchases large number of mortgages from lenders for purposes of pooling them into securities that can be invested in by the public. To bolster transparency and liquidity in the investment market, Fannie Mae provides performance data (monthly payments, payoffs, defaults, foreclosures, etc.) for mortgage loans in their portfolio that were originated after 1999. The data they provide is limited to single family, fixed-rate mortgages with maturities of 30 years or less.

The current dataset contains performance data on over 35 million loans and has over 50 data points (features) collected on each loan. Fannie Mae provides these files on a quarterly basis depending on the quarter in which they purchased the loan from lenders. I chose datasets from the first two quarters of 2012 for the exercise. I wanted a dataset that reflected changes enacted after the financial crisis, but that also had lengthy period in which defaults may occur.

Fannie Mae provides data for each origination quarter in two files. The first file provides origination characteristics of the loan range from borrower and collateral information (i.e. number of borrowers, the type of home, whether this is the borrower's first purchase) , terms of the loan (i.e. maturity, interest rate, original balance), and metrics used to evaluate credit worthiness (i.e. borrower's credit score, ratio of size of loan to value of home, the ratio of the borrower's debt to their income). The origination files for the first two quarters of 2012 had approximately 600 thousand loans each. To assess historic performance of loans, Fannie Mae also provides a file containing the status of the loan (i.e. current balance, age of the loan, whether the loan has defaulted or foreclosed upon) at each monthly snapshot over its life to date. The monthly performance files for the first two quarters of 2012 were each nearly 3GB in size containing tens of millions of monthly loan snapshots.

For this exercise, I chose to use the dataset from the first quarter of 2012 to mimic a form of "existing" knowledge before validating the quality of the models on the dataset from the second quarter of 2012. Data link (requires creation of login):
https://loanperformancedata.fanniemae.com/lppub/index.html

Process Overview

Given the large size of these datasets, I first built a notebook that would augment the origination characteristics file with a simple flag to indicate if the loan has experienced difficulty at any point in its life. Specifically, this flag indicates whether the borrower had ever fallen more than 90 days behind on their payments. I removed loans where there were status gaps in the monthly snapshots. I also removed loans that were missing key features, like credit score, debt-to-income, loan-to-value, etc. While the original files are not provided due to their size, I am including the datasets I created. This notebook is titled "FNMA - Dataset Creation". This code will not execute given the original files are not provided.

I then performed an initial exploratory data analysis on the 2012-Q1 dataset. This dataset had 637 thousand loans, and less than 1% of those loans had experienced a default. Due to the computational challenges exploring this large a dataset, I performed analysis on a random sample of 10% of the entire dataset. My first findings confirmed that this sample had a default rate within 1/10 of a basis point of the entire sample. Analysis on the sample set revealed some obvious and interesting facts. The distribution of credit scores was in fact lower for loans that defaulted. Defaulting loans also had higher density of DTI values than non-defaulting loans. Interestingly, the

defaulting loans saw a higher proportion of loans issued for the purpose of taking cash out of a home (cash-out refinance). Defaulting loans were also more likely to have just one borrower. Visualizations are contained in the "Exhibits" section below. They were created with matplolib and seaborn. This notebook is titled "FNMA - EDA".

I then proceeded to the modeling phase. My goal was to try a number of different modeling types as well as various parameters for each. Models I included were Decision Tree, Random Forest, Multi-Layer Perceptron, Support Vector Machine, and Logistic Regression. Given the processing consumed with the SVC library from sklearn, I chose to use the linear optimized LinearSVC library. The models were fitted, and optimal parameters discovered, using pipeline and grid search techniques. Models were fitted on a 10% sample of the 2012 Q1 dataset. This sample size provided a higher quality of fit, while allowing for a manageable amount of processing time. Once models were fit and tested on the 2012 Q1 dataset, I performed a validation of the models using the 2012 Q2 dataset. I performed analysis to ascertain key features used in making predictions. Finally, I performed analysis to evaluate the predictions being made by models at an individual sample level. This notebook is titled "FNMA - Modeling".

Model Discussion

My modeling results were mixed, with several performers (Decision Tree, Random Forest, and MLP) offering little to no predictive power with the parameters and pipelines executed. These models simply chose the non-defaulting loan classifier due to the imbalance with the target. This was despite setting the scoring parameter to the GridSearchCV function set for "recall." In sklearn's documentation setting the scoring parameter to "recall" in a binary target situation as this should seek to find the highest recall of the "positive" class (i.e. defaulting loan). This aligns with the goal I had of identifying all loans that will default. A simple overall accuracy would be insufficient in this instance. The highest overall accuracy any model could likely achieve would be to predict no borrower will default.

Clearly, the two highest quality models were built on Logistic Regression and Support Vector Classifiers. Interestingly, the quality of these models highlighted an interesting trade-off between recall of the default classification versus recall of the non-default classification. The Logistic Regression model identified almost 90% of the defaults in the 2012 Q2 dataset. However, this was achieved at the expense of the model's ability to recall the non-defaulting class. In fact, the Logistic Regression model predicted nearly as many defaults as it did non-defaults, which seems a bit inappropriate given the default rate we saw during EDA. The Support Vector Classifier model identified just 75% of the defaults, but achieved that in a far less aggressive manner. These results can be seen in the ROC curves in the "Exhibits" section below. It is interesting to note that the results received on the testing set from the 2012 Q1 dataset were similar to validation performed on the entire 2012 Q2 dataset.

A quick study of the feature coefficients coming out of the models highlights some important, though somewhat obvious, findings. Features such as credit score, debt-to-income, loan-to-value were all prominent as expected. Interestingly, the number of borrowers and the investor occupancy status were prominent factors in the more sensitive Support Vector Classifier model. This corresponds to findings from EDA as mentioned earlier. The top 10 features for predictions of the defaulting and non-defaulting classes for both Linear Regression and Support Vector Classifier can be seen in the "Exhibits" section below.

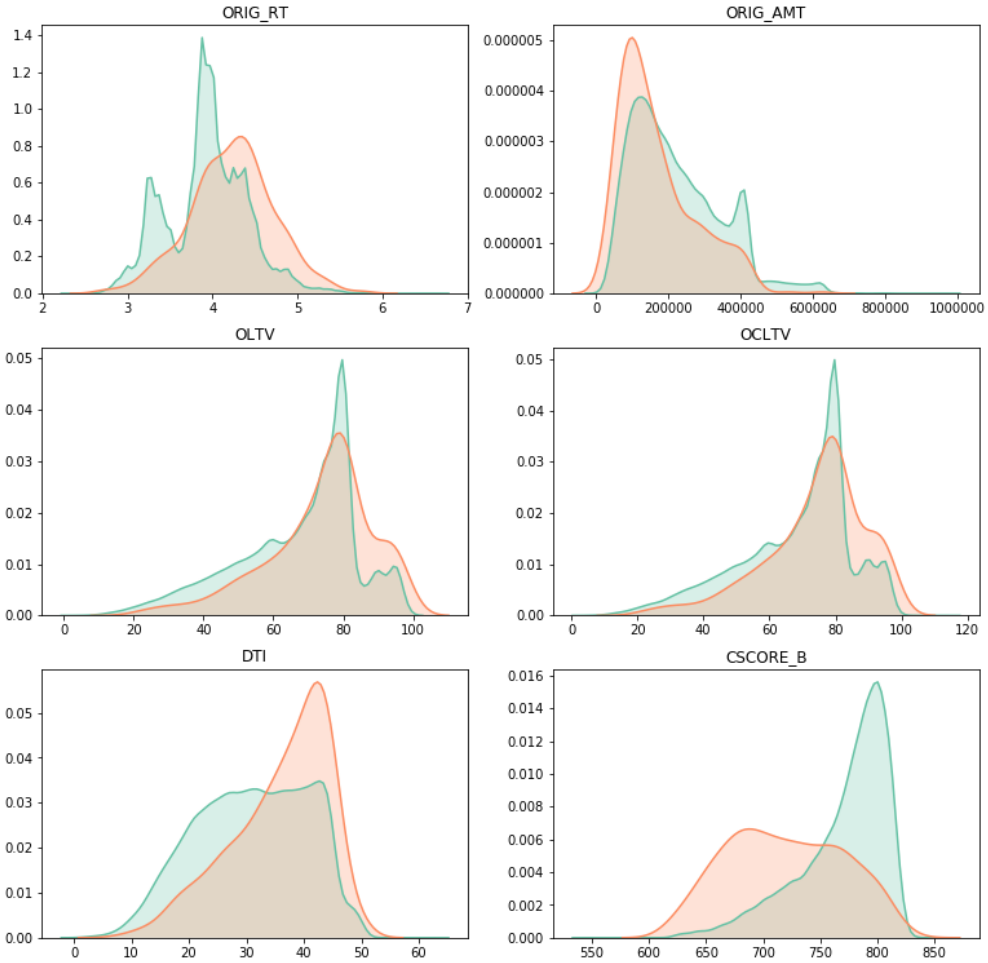
Finally, Logistic Regression provides very useful functionality I wanted to explore to analyze how likely a loan was to default, based on a model. These predicted probabilities serve as a mechanism for providing some gradation between loans. This is useful as not all loans should receive greater scrutiny and monitoring than others. For example, a bank may use these probabilities and their own custom threshold to determine whether or not to grant a loan. Some of my preliminary modeling (not provided) gave some loans a 97%+ probability of default. Intuitively, it does not make sense to say that someone who received a loan has a 97% of defaulting, based on the relatively limited amount of data that can be captured during an underwriting process. Someone could have a 600 credit score, a high LTV and DTI, but have just won the lottery or received a large inheritance. Fortunately, I had more realistic values coming out of my final Logistic Regression model. The highest predicted probability of default was just 73%. The lowest was nearly 30%, which feels a bit high, though maybe not unreasonable.

Conclusions

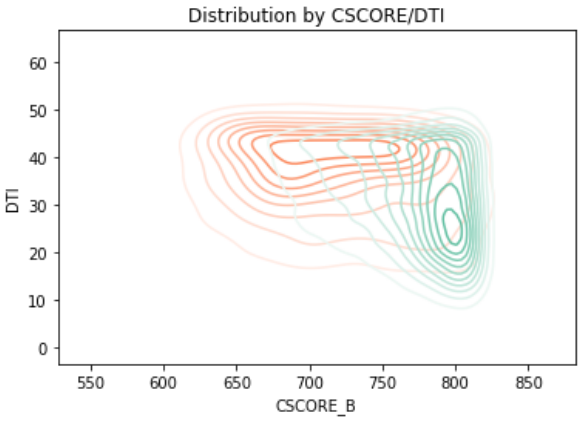
The results from the Logistic Regression model highlight why it has been so widely used in mortgage loan default prediction. The results from the Support Vector Classifier present an interesting alternative. I believe with more time, and more processing power, all models have room for dramatic improvement. The large amount of samples, coupled with the severe imbalance of the target, present a difficult challenge.

Exhibits

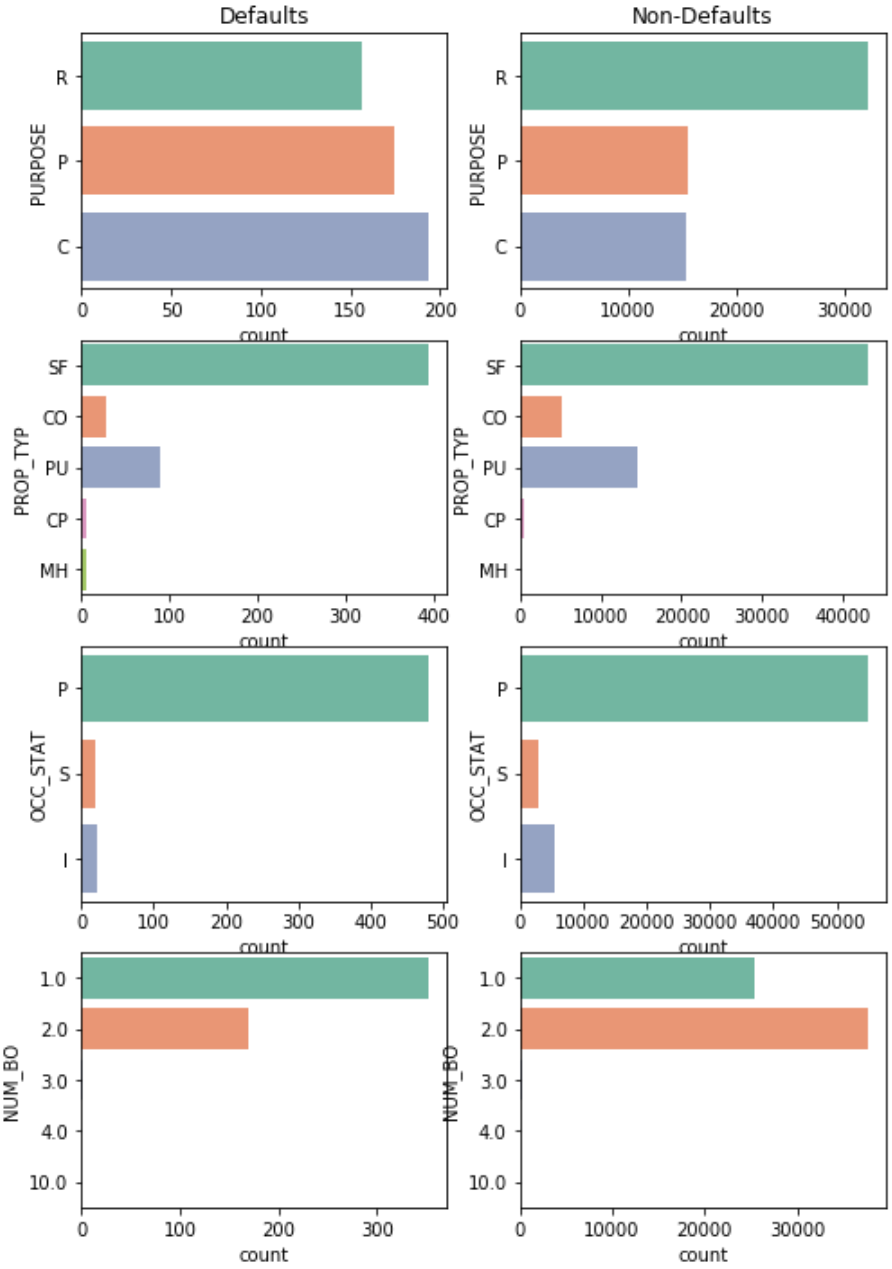
EDA - Density Distributions for Numeric Features (Green=Non-Defaults; Orange=Defaults)



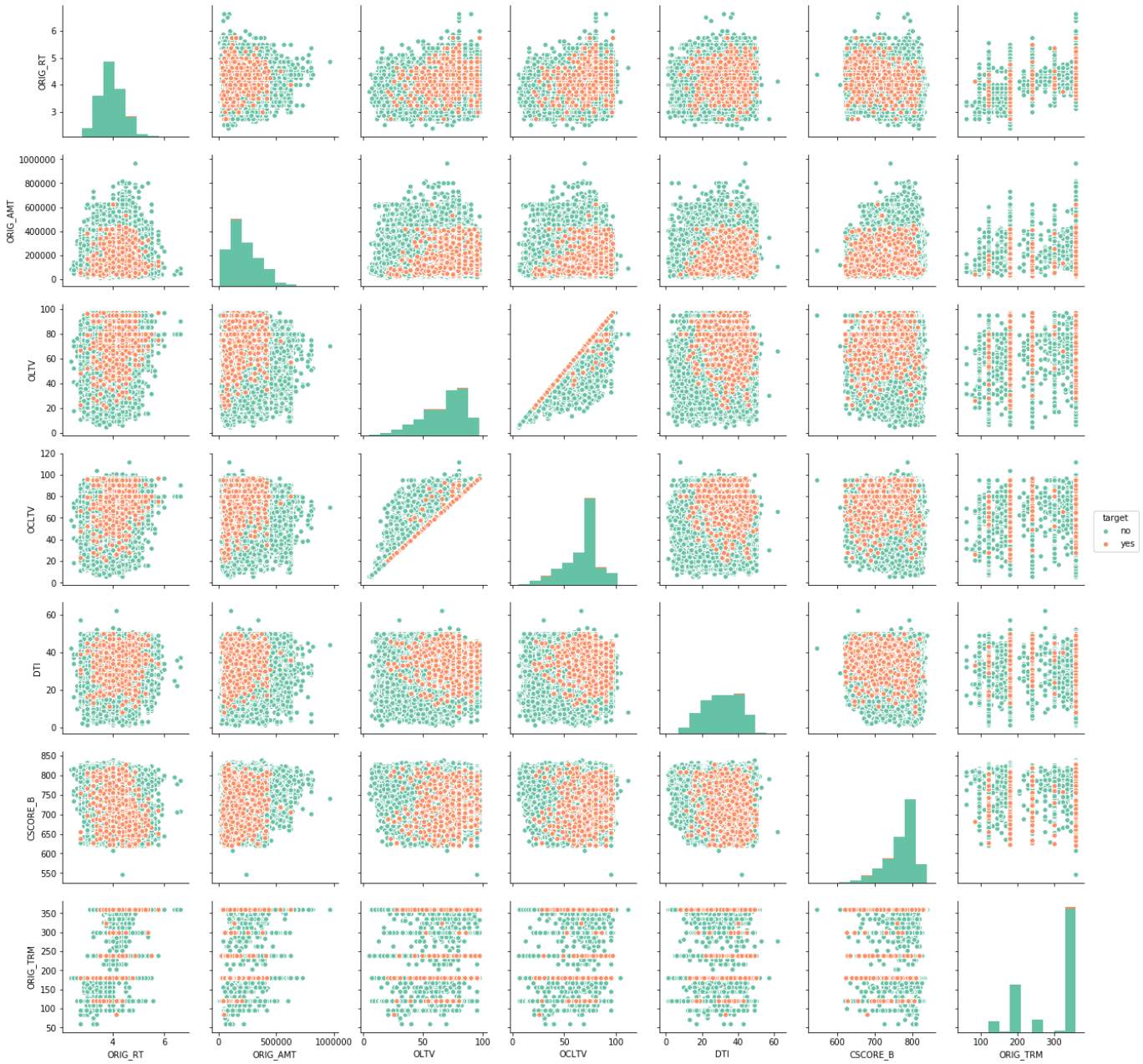
EDA - Density Distribution on Two Features (Green=Non-Defaults; Orange=Defaults)



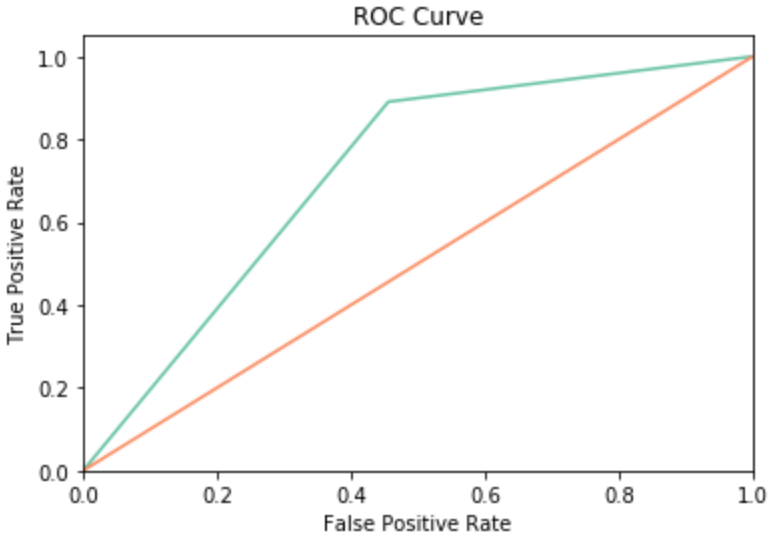
EDA - Value Distributions for Select Categorical Features



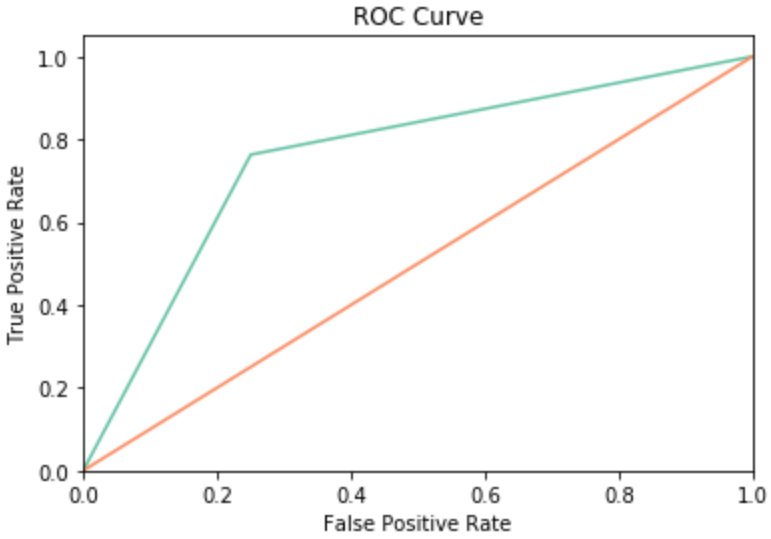
EDA - Correlation Matrix (Green=Non-Defaults; Orange=Defaults)



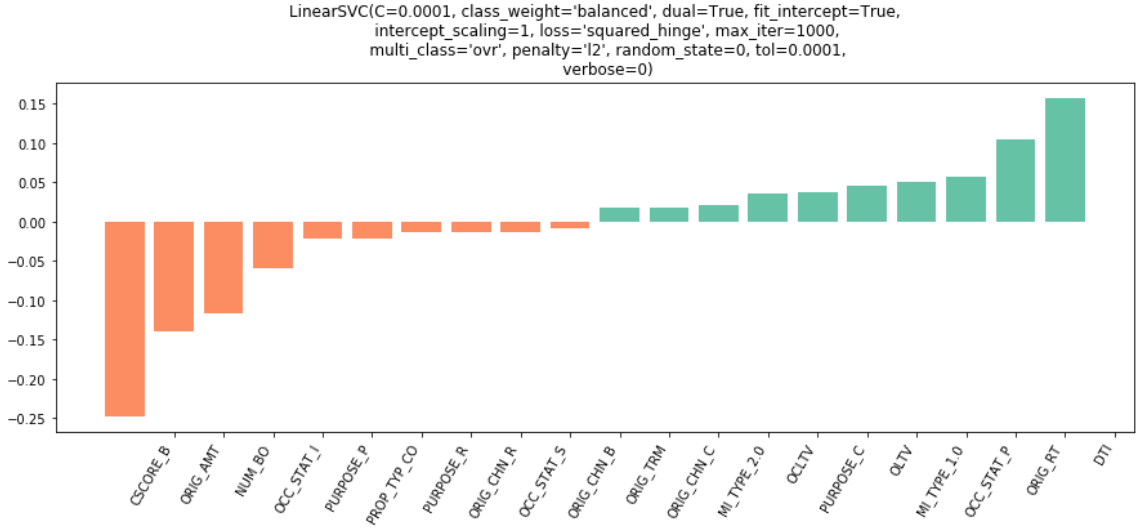
ROC Curve - Logistic Regression



ROC Curve - Support Vector Classifier



Feature Importance (Green=Non-Defaults; Orange=Defaults)



Predicted Probabilities by Feature (Green=Non-Defaults; Orange=Defaults)