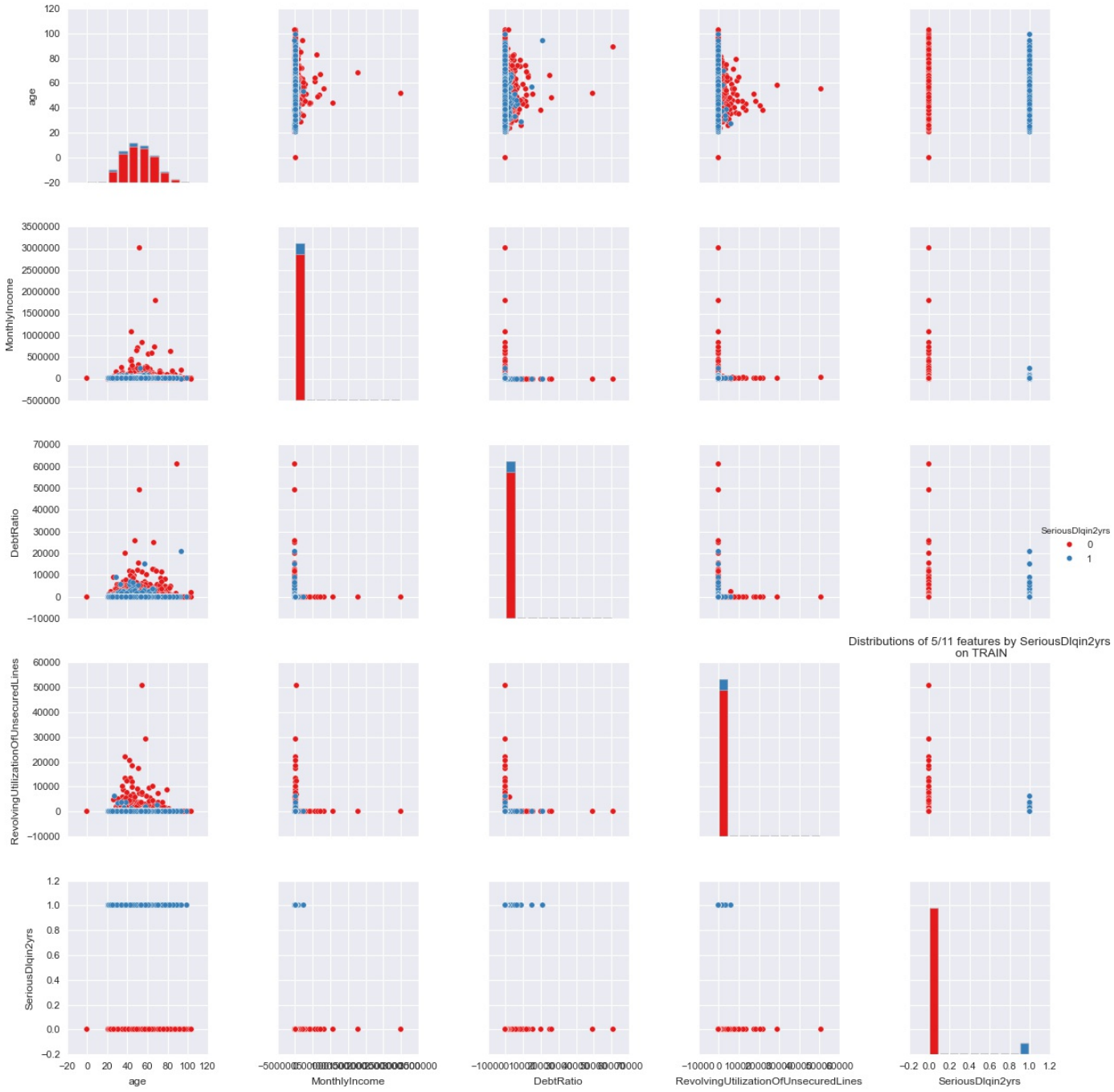
Pipeline

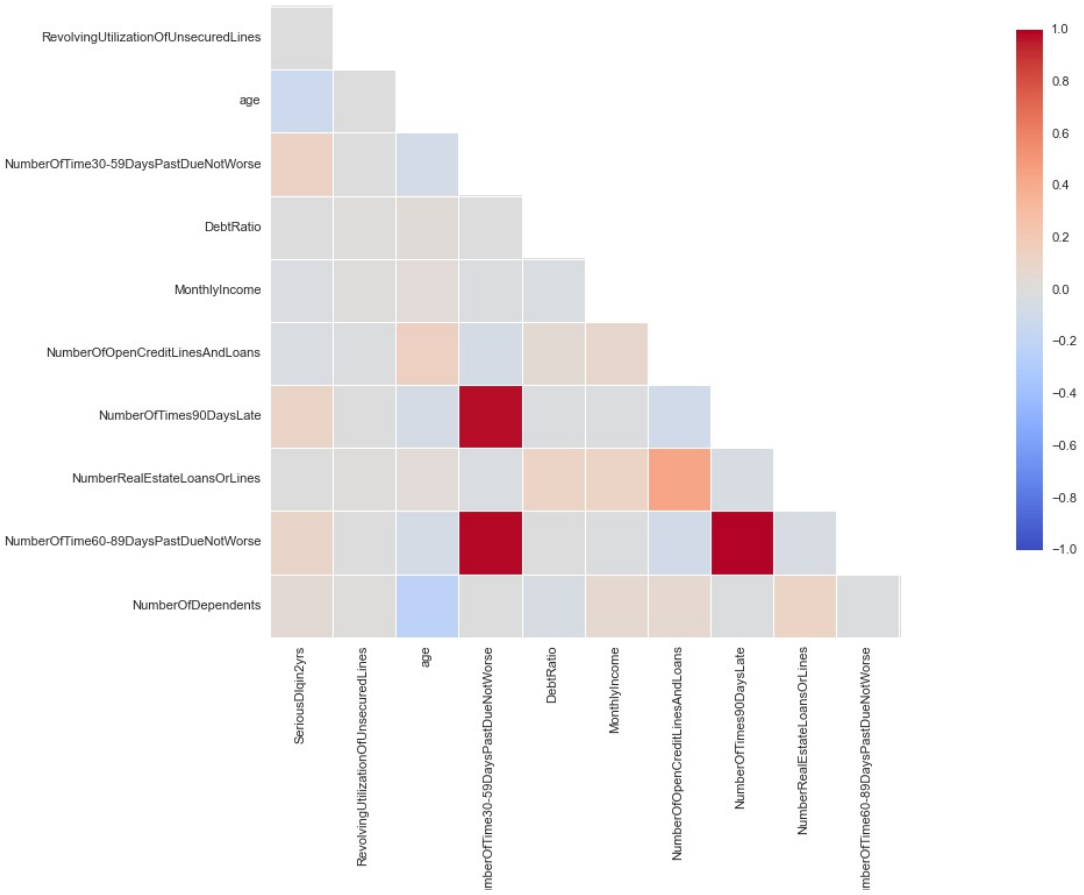
The purpose of this project is to predict which people will experience financial distress in the future. To accomplish this goal, I trained a Random Forest classifier on a dataset that includes key features such as a person's monthly income, debt ratio, number of dependents, and number of open credit lines and loans. The overall strategy to fitting a good predictive model to this dataset include: understanding the dataset via exploratory data analysis, imputing missing values, transforming features, learning and validating the model for a particular model class, and finally predicting which persons in the test set will most likely experience financial distress in the future.

My first task was to explore the dataset in order to gain insight into the problem. One straightforward way to investigate the data is by generating a series of pair-plots and look at the pair-wise relationships between the features.

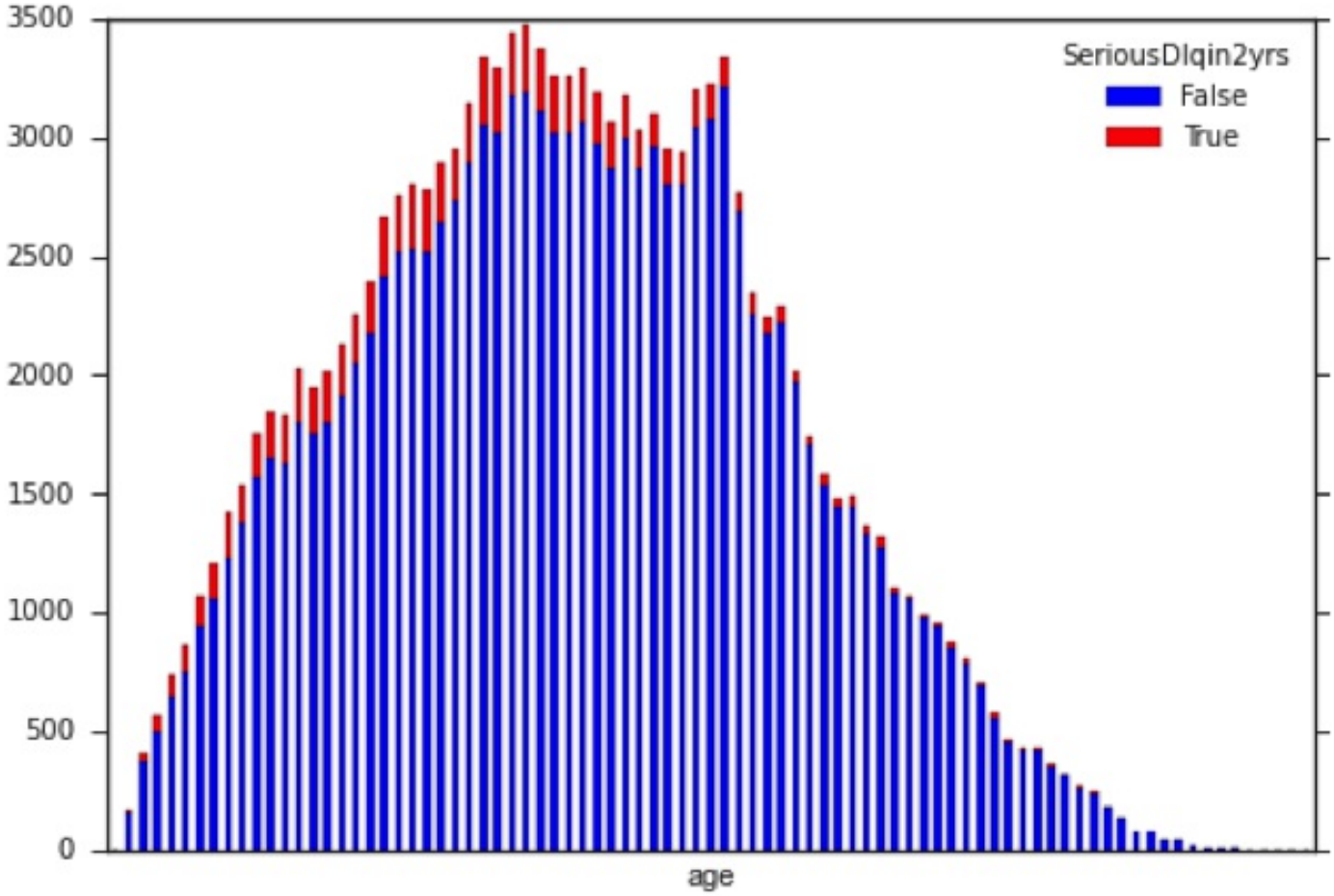


I also inspected more specific questions about the dataset to look for potential redundancies in predictive power. Which features might measure one another but fail to explain financial? The correlation plot below shows that the features measuring 'Days-late-to-payments' are correlated together, mostly *along* financial distress, rather than with it.

TRAIN_PARENT
(Anti)Correlation Matrix
Index([u'SeriousDlqin2yrs', u'RevolvingUtilizationOfUnsecuredLines', u'age',
u'NumberOfTime30-59DaysPastDueNotWorse', u'DebtRatio', u'MonthlyIncome',
u'NumberOfOpenCreditLinesAndLoans', u'NumberOfTimes90DaysLate',
u'NumberRealEstateLoansOrLines',
u'NumberOfTime60-89DaysPastDueNotWorse', u'NumberOfDependents'],
dtype='object')
(150000, 11)



Or for example, is there a relationship between age and financial distress? Do younger people experience financial distress at a higher rate than older people? The graph below suggests that there is a sharp drop in serious delinquency as age increases, somewhere following the middle-aged range.



After exploring the dataset, the first challenge in this project was to impute missing values for various features, the most important missing feature being monthly income. Also among the data were mistakes and encoded missings, mascarading as real numbers such as codes ‘96’ and ‘98’ (a commonly used in survey questionnaires for “DON’T KNOW”, ”N/A”, or “REFUSED”). Decoding these values and errors to missings clarfies the different relationships between missing data monthly income and their nonmissings counterfactuals. Number of dependents, age, and monthly income are more closesly related when missing than when present. Since missing values may not be missing completely at random, I estimated imputation model.

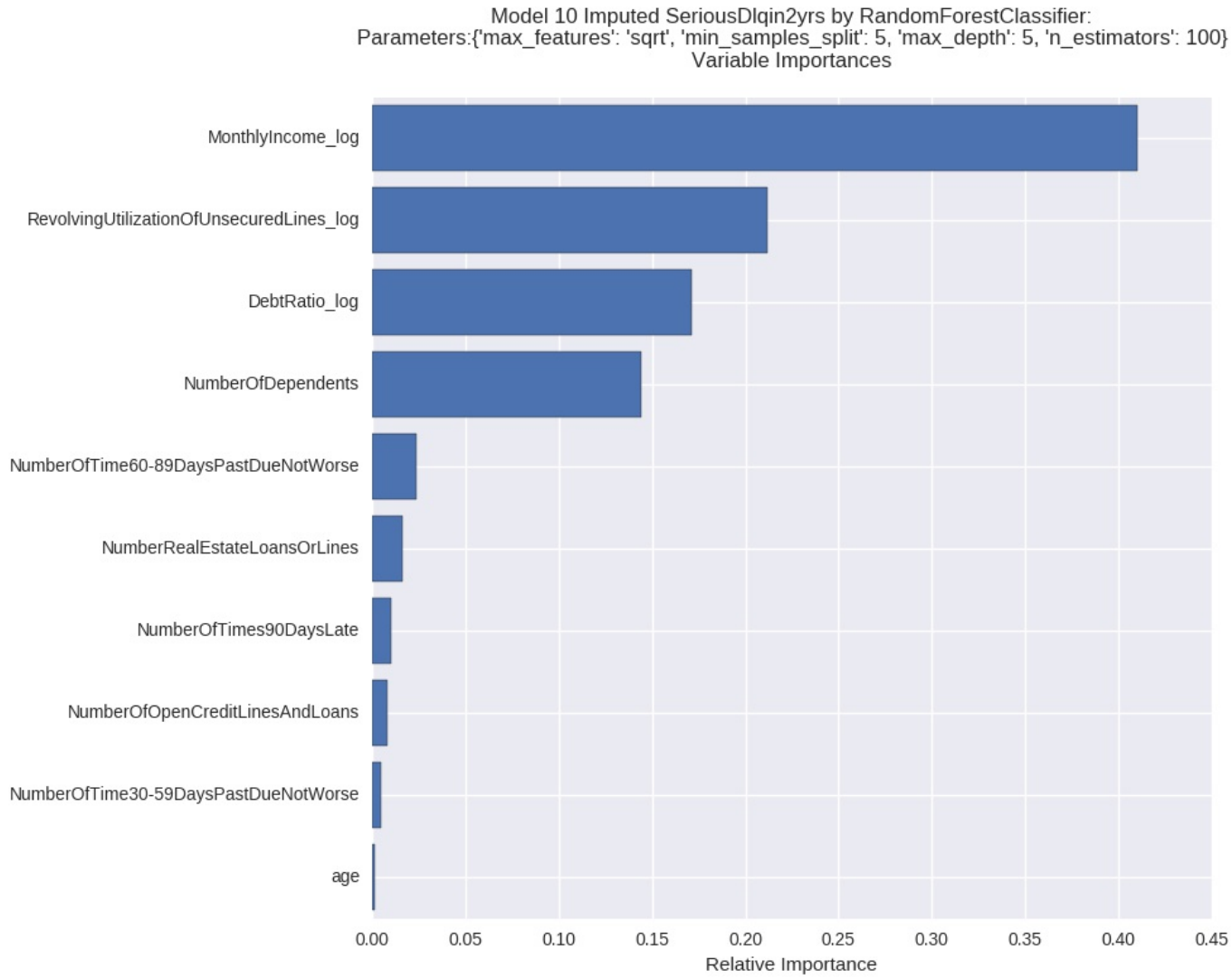
There are a number of approaches one could take when imputing the missing monthly income values, such as imputing the mean or the median. Instead, I chose to predict monthly incomes by fitting a regression model to the training data. The advantage of predicting the monthly income over imputing mean or median values is that a regression model has a higher likelihood of capturing the true variance within the dataset.

This time, I cross-validated multiple-imputation in tandem with the classifier for prediction of ‘Serious Financial Delinquency’, such that the classification and the imputation learned on the same training and validation sets. Instead of first validating for Monthly Income and then another for ‘Number of Dependents’ separately, all imputations learned on subsets of the training and validation models for serious financial distress.

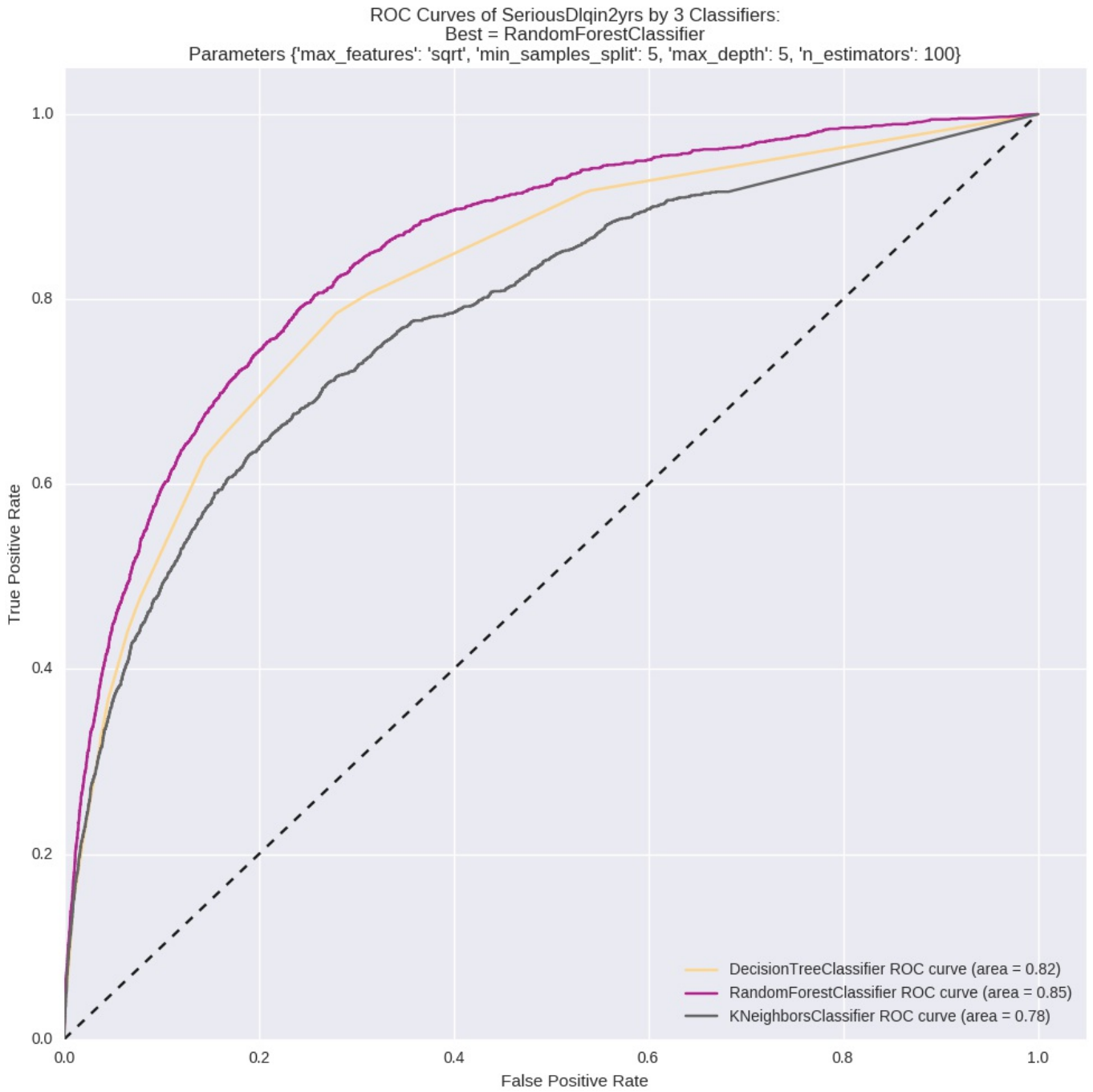
I also transformed ‘Monthly income’, 'DebtRatio', and ‘Revolving Utilization Of Unsecured Lines’ to log space during the cross validation. I did this to control for the large range of monetary values

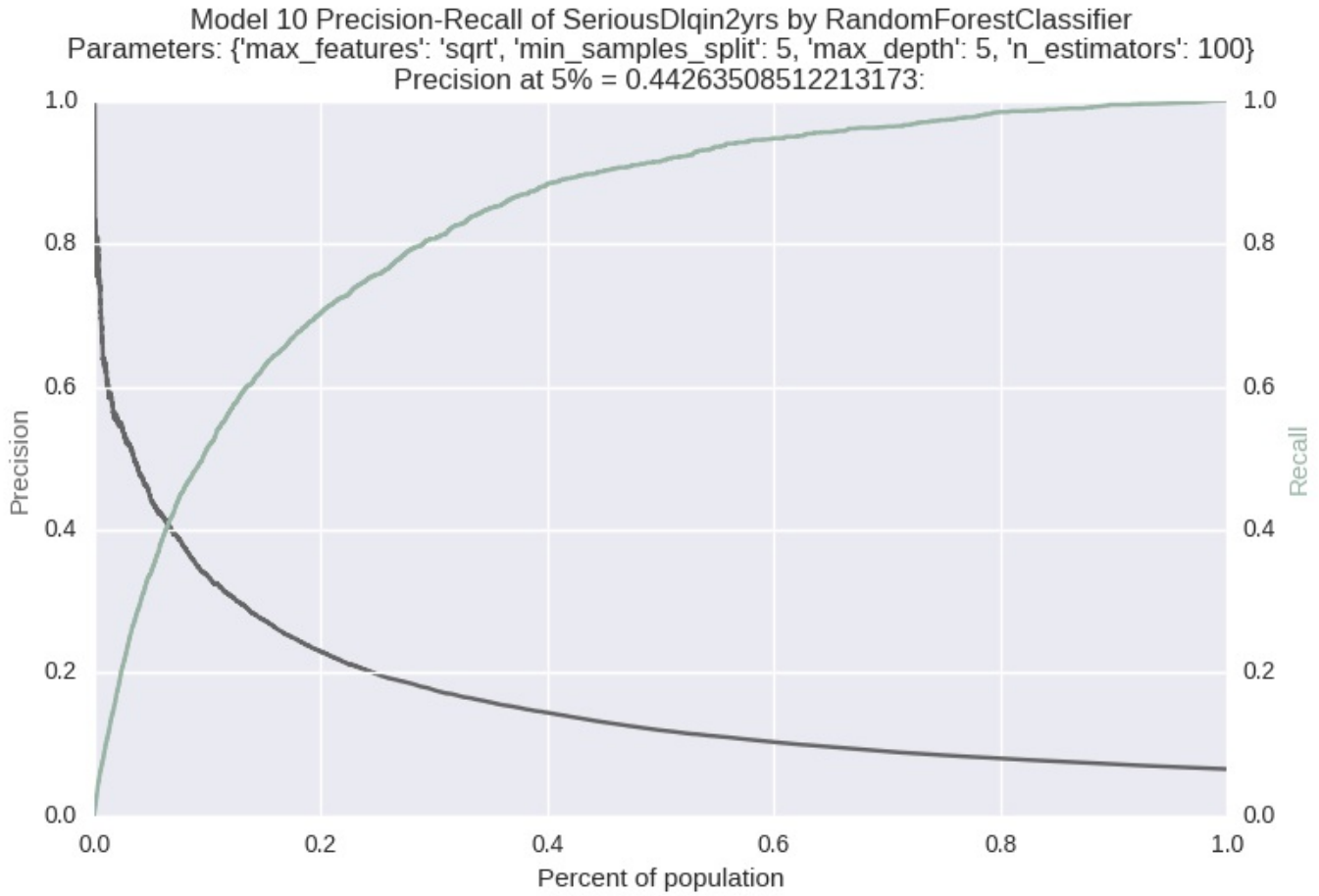
that are present in the dataset. In addition, this transformation to monthly income modified its distribution to something that was qualitatively Gaussian.

In the last step of this project I trained a Random Forest Classifier, a Decision Tree Classifier and a KNN Classifier on the now complete training dataset. KNN was the slowest to train, DT the quickest. The Random Forest Classifier was chosen based on ROC_AUC scores constrained to a threshold of.05 on the validation set. I observed that the most important variables for predicting financial distress were ‘Monthly income’, 'DebtRatio', and ‘Revolving Utilization Of Unsecured Lines’.

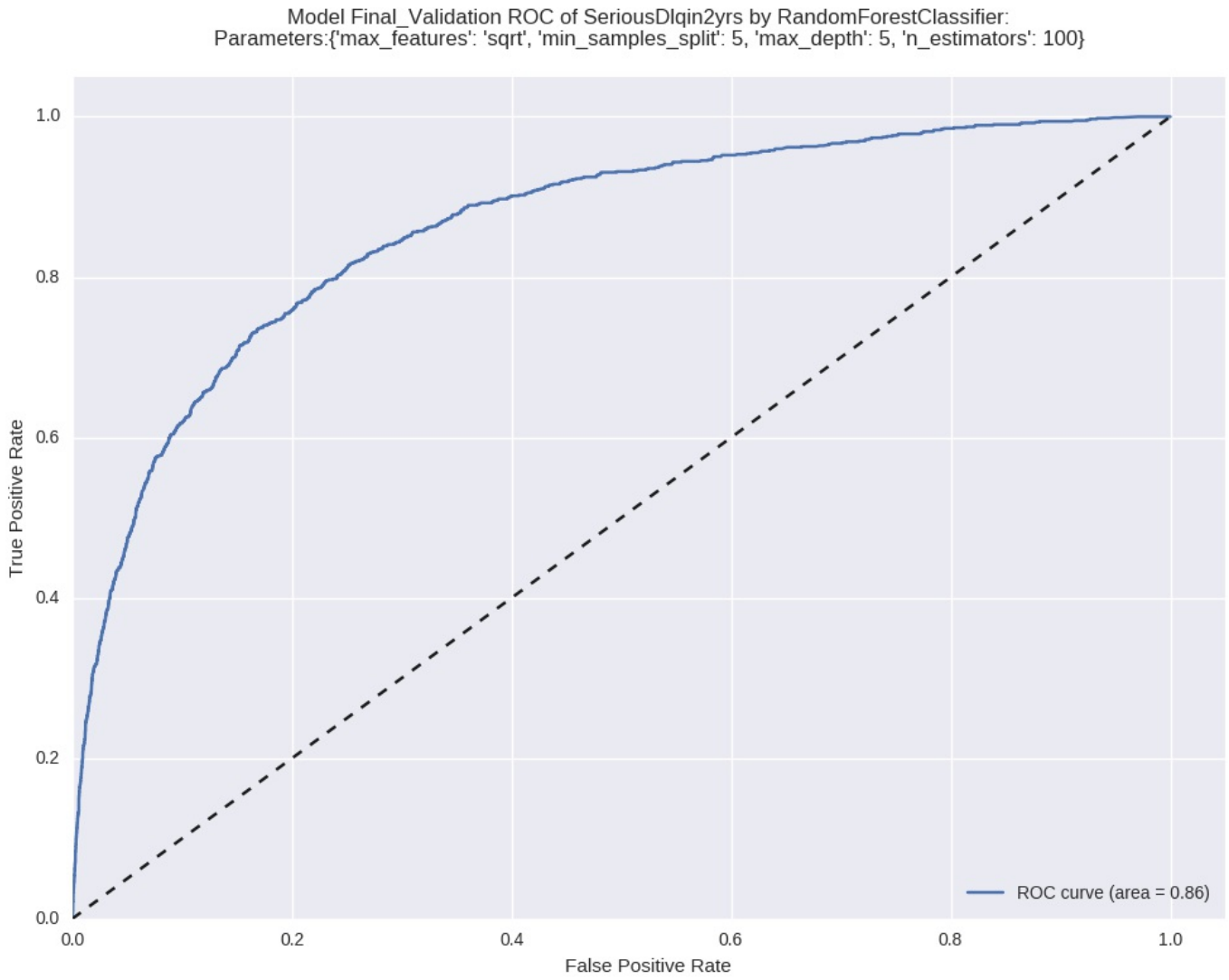


You can see a summary of how well this classifier worked by looking at the ROC Curve and precision-recall curve below





These metrics were further validated on a holdout set, which was imputed with the statistics cross-validated on the training set.



The ROC_AUC score of the RandomForestClassifier is increased from 85% to 86%. This result suggests the model did not overfit to the training data.