**Predicting Default Risk**

Project Overview

We are a loan officer at a young and small bank (in operation for two years) that needs to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not. We’ll use a series of classification models to figure out the best model and provide a list of creditworthy customers to our manager.

The Business Problem

We work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Our team typically gets 200 loan applications per week and approves them by hand.

Due to a financial scandal that hit a competitive bank last week, we suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process this week!

Our manager sees this new influx as a great opportunity and wants us to figure out how to process all of these loan applications within one week.

For this project, we will analyze the business problem using the Problem Solving Framework and provide a list of creditworthy customers to our manager in the next two days.

We have the following information to work with:

1. Data on all past applications
2. The list of customers that need to be processed in the next few days

Step 1: Business and Data Understanding

Our project should include a description of the key business decisions that need to be made

Step 2: Explore and Cleanup the Data

To properly build the model and select predictor variables, we need to explore and cleanup our data.

Here are some guidelines to help clean up the data:

1. Are any of our numerical data fields highly-correlated with each other? The correlation should be at least .70 to be considered “high”
2. Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
3. Are there only a few values in each subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability.
4. The clean data set should have 13 columns where the Average of [Age Years] should be 36 (rounded up)

Note: If we are deciding to impute any data field, for the sake of consistency in the data cleanup process, impute the data using the median of the entire data field

Step 3: Train our Classification Models

We should choose 70% to create the Estimation set and 30% to create the Validation set. Set the Random Seed to 1 if we are using Alteryx.

Train our dataset using these models:

* Logistic Regression
* Decision Tree
* Forest Model
* Boosted Tree

Step 4: Writeup

Compare all of the models’ performance against each other. Decide on the best model and score our new customers.

Important: Our manager only cares about how accurate you can identify people who qualify and do not qualify for loans in this problem