The Business Problem

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your 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, you 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!

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

Fortunately for you, you just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants.

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

You 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

Steps to Success

Step 1: Business and Data Understanding

Your 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, you need to explore and cleanup your data.

Here are some guidelines to help you clean up the data:

1. Are any of your 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 a 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. Refer to the "Tips" section to find examples of data fields with low-variability.

4. Your clean data set should have 13 columns where the Average of `Age Years` should be 36 (rounded up)
   Note: If you decide 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 your Classification Models

You should choose 70% to create the Estimation set and 30% to create the Validation set. Set the Random Seed to 1 if you're using Alteryx.

Train your 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 your new customers.

Important: Your manager only cares about how accurate you can identify people who qualify and do not qualify for loans for this problem.
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan.

Data

credit-data-training.xlsx - This file contains all credit approvals from your past loan applicants the bank has ever completed.
customers-to-score.xlsx - This is the new set of customers that you need to score on the classification model you will create.