Every time someone lends something of value to another person, there is always a risk that the item will not be returned or will be returned in worse condition than when it was lent. For example, a car rental company can rent a car and it can be returned with damage, or a library can lend a book and that person will never return the book. This is the same concept for financial institutions lending money to consumers. When a bank lends money to someone to buy a house, they are lending hundreds of thousands, sometimes millions of dollars and they expect to get that money back over time. With the very large amount of money involved in these transactions, it's important for the bank to know if they are going to get their money back before lending it out.

In this project, I have taken data provided to the Kaggle community by Home Credit Group and attempted to predict whether a lender will default based on data gathered by the company. When exploring this data, I found that it was very dirty. There are 122 columns in the application_train.csv dataset and many of them have a large percentage of null values. As a result, most of my time was spent on cleaning the data and selecting the necessary variables before I could train a model to predict the default value.

For the heavy lifting in this project, I utilized R and many packages within R. Missing values were imputed using the mice package and the dummies package was used for one-hot encoding. Many models were tested for predicting the target variable including a random forest model from the randomForest package, naïve Bayes from the e1071 package and a k-nearest neighbors model from the class package.

References

Home Credit Group & Kaggle.com. (2018). Home Credit Default Risk. Raw data. Retrieved from https://www.kaggle.com/c/home-credit-default-risk/data

Sanguansintukul, Siripun (2018) Class Materials from MSDS 680 and MSDS 660