# FIN 580: Final Project
# EXECUTIVE SUMMARY

Due: Upload to Compass by 11:59pm on Friday, December 7, 2018

Your Team Name (pick one): __Wonders_____

Select whether this is an individual or group submission. No more than 3 members per group. Beyond the fact that all group members may submit the same answers, each submission must be separate work.

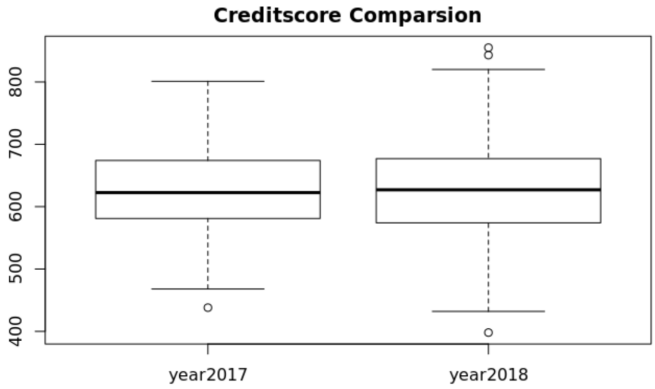
☐ Individual Submission
√ Group Submission. List group member names: _Jinran Yang, Wanting Yao, Tianlun Zhao _____

## Case Overview

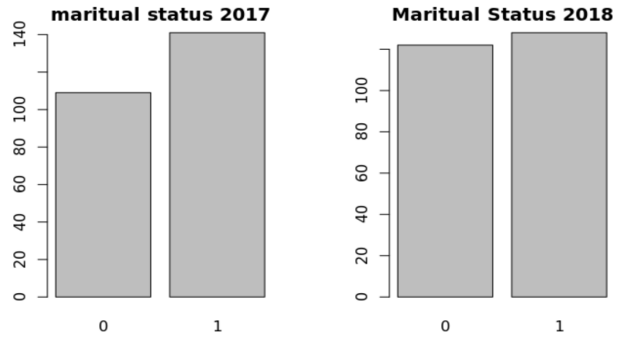
Chess Bank wants to make maximum profit through løan in 2018. Profit is defined as the difference between the amount of money paid back by the applicants (short for "amount paid") and the amount of money lend by the bank to loan applicants (short for "loan amount"). The purpose of this project is to accurately predict those who will make loan repayment (include interest) more than loan amount. We need to build statistical models based on data in 2017 and select appropriate models to predict the amount paid of the applicants in 2018, and based on the predictions, make decisions about giving them loan or not to achieve the maximum profit.

## Methodology

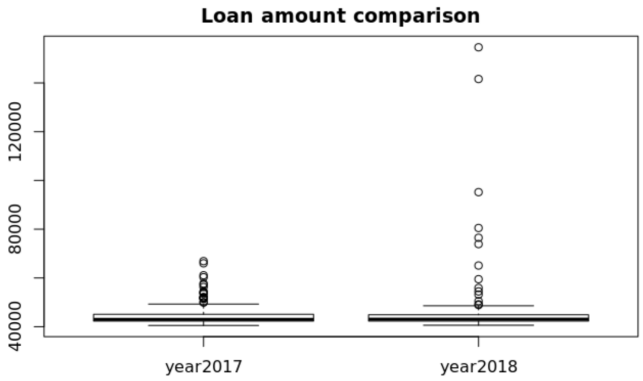
Before applying models to data in 2018, we need check whether the distribution of 2018 and that of 2017 are similar. Due to the space limitation, we only show partially result. Based on the graphed boxplot below, we are comfortable to assume that the data in 2017 and 2018 follow a similar distribution. Thus, the model created using 2017 data can be applied to predict data in 2018.




Picture1: credit score distribution comparison




Picture 2: marital status comparison




Picture 3: Loan amount comparison

- **Step 1: Clean and process data in 2017**

First, we check whether there are missing values and duplicated records in the dataset. After guaranteeing the dataset is clean, we transformed the categorical variables - such as date, statecode, education, marry status, taxdependent - into factors, preparing them for further analysis.

- **Step2: Select variables related to amount paid**

Second, since our goal is to fit a model to explain the amount paid, we check the relationship between each variable and amount paid one by one by fitting linear regression model. For loan applicant's wage income 1 & 2 years ago, those two variables are highly related, so we take average of them to create a new variable, W2INC, to stand for the income status of loan applicants. In these linear regression models, we selected those related variables using p-value. The variables we include are: credit score, W2INC, married status, tax dependent, asset, debt, average home price.

- **Step 3: Build models**

We tried different kinds of models: linear regression, LASSO and tree-based model (Decision Tree, Random Forest, and Boosting). We spilt the 2017 data (250 in total) into train data (220) and test data (30). And we train our models using train dataset and using the model we trained to predict the amount paid in the test data and then calculate the accuracy. Since we believe that models with better performance on predicting the data in 2017 should perform better on data in 2018.

- **Step 4: Select the best model and apply it to data in 2018**

To compare the performance of each model, we check the models' accuracy of loan prediction in 2017. Based on the accuracy (Table 1), these five models have rather similar accurate rate but none of them has a very high accuracy rate. So we choose to apply these 5 models to data in 2018. If more than 3 models' prediction result is giving loan, then we will decide to give loan. Otherwise, we are not going to give them loan.

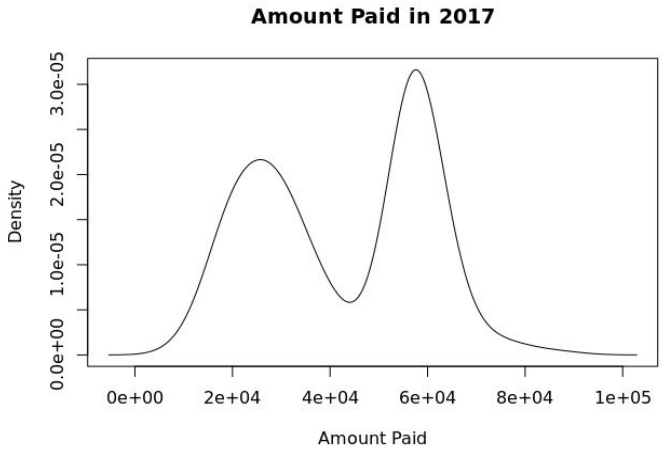
Table 1: Accuracy Check Result of Different Models

| **Models** | Linear Regression | LASSO | Decision Tree | Random Forest | Boosting |
|---|---|---|---|---|---|
| **Accuracy (%)** | 0.6666667 | 0.70 | 0.60 | 0.7333333 | 0.7333333 |

# Conclusion

Based on the prediction in 2018, Chess Bank should give loan to selected 123 people. The total estimated loan amount is $5,460,000, and the estimated profit is $629,064.90.

As the prediction result of 2017 shows, the accuracy is not very satisfying. So we further look at the distribution of amount paid in 2017. The density plot is as follows:




Picture 4: Amount paid density distribution in 2017

As the above figure shows, there are two peaks which indicates that the amount paid follows a multimodal distribution. Thus, it is more appropriate to build two models, the first one is the amount paid under $40,000 and the other is the amount paid above $40,000. However, we need more data to build the two models. Due to the limited amount of data given, we are not sure that building two models would accurately reflect the general trends of underlying dataset. This approach should be useful if sufficiently more data were given to train the models.

One more thing to note, we assume data in 2017 and 2018 follow similar distributions. Upon closer examination, we notice that there are several outliers in the 2018 data which could cause a violation of our earlier assumption. Because of the size of the dataset, We don't know if these are just noise or it could potential cause skewness in the distribution. If we were able to obtain a larger dataset, we would be able to take a closer examination with the issue and offer more informative results.