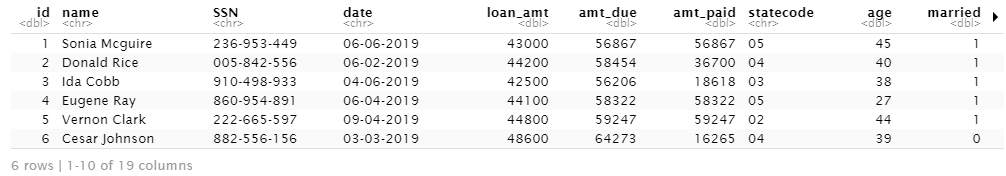
# Case Overview

Chess bank recently received 250 loan applications and wanted to decide who to give loans so that the profit would be maximized. As Chess’s chief data analyst, I will need to analyze historical loan applications and their payment information to learn the relationship between individual characteristics and loan payment behaviors and to predict the payments for 2020 applicants. Then, I could increase the chance not only to get the money back on time but also to get the interest payments to maximize the profit with the predicted results.

# Methodology

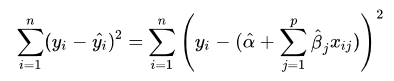
**Preprocessing the Data**

By importing the 2019 loan applications data, I explored the head of the dataset to understand the general look on the dataset.

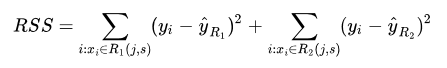
I also explored the variable names to check if it matches the given ones, and I checked for duplicated loans and if the number of loans matches, where they all match. Then I cleaned the data by remove missing values and select variables that I will need to predict the loan amount that will be paid: I did not include four variables: “id”, “name”, “ssn”, “date” that are very individual and unlikely to relate to the amount paid variable. I prepared the date by splitting it into a train set and a test set by 0.8 : 0.2 and set the seed to ensure that I get the same result if started with the same seed.

**Applying Predictive Models – Linear Regression, Lasso, and Decision Trees**

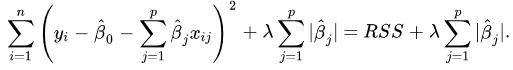
I chose multivariate linear regression, decision trees and random forest (including bagging and boosting) as the predictive models. Multivariate linear regression was used when there are more than one explanatory variables, and the function below was used to solve for 1 + p coefficients: α and β that minimize residual sum of squares.



Tree-based methods involve stratifying or segmenting the predictor space into different regions. The prediction of the outcome variable for a given observation is typically equal to the mean of the training observations in the region where it belongs. The goal is to define R(j, s) such that it minimizes the residual sum squares as below:

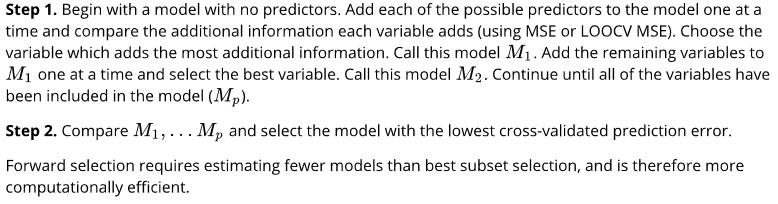


Lasso is a technique for constraining coefficient estimates to avoid fitting the models to random patterns in the data by setting coefficients **λ** based on shrinkage or selection techniques.

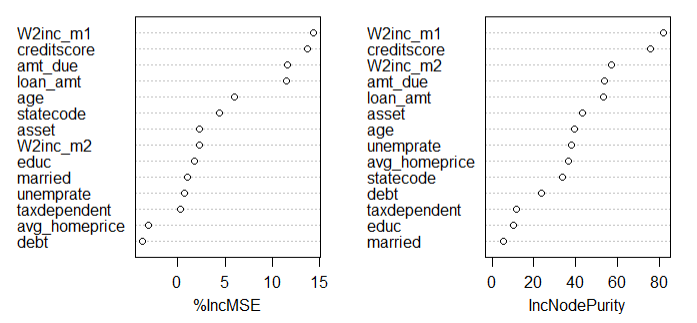


**Improving the Linear Regression Model –Cross Validation and Variable Selection**

There were total 14 variables that I chose to analyze the relationship: credit score, loan amount, asset, state, average home price, age, unemployment rate, educ, debt, tax dependent, married, amount due, W2inc\_m1: applicant’s income 1 year ago, W2inc\_m2: applicant’s income 2 year ago. While additional variables might improve the predictive power of the model, it may also lead to lower mean squared error and overfitting problems. Cross-validation is a resampling method that will help overfitting the model on each sample. I used Leave-One-Out Cross Validation that compares the mean squared errors for each estimate, which leaves one observation out each time in the training data set. Then I select important features out of 14 variables that have least mean squared errors after cross-validation using Forward Stepwise Selection steps as below:



The result formula after variables selection is amt\_paid ~ W2inc\_m1 + creditscore + statecode + age + amt\_due, which is also similar to the top important features shown by random forests:

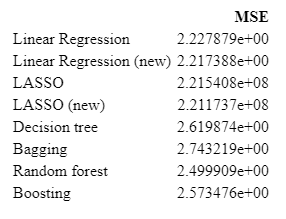


**Improving the Tree Model –Bagging, Random Forests, Boosting**

Bagging, Random Forests, Boosting are aggregate multiple trees that are designed to reduce the variance of decision trees. Bagging is taking the average of the estimates from each tree built in each sample from the data. Random Forests build decision trees based on bagging trees but with a feature that forces the trees to be different. Boosting is building trees with the information from previous trees, which build trees slowly to avoid overfitting the data.

# Conclusion

To conclude, I compared 8 models: Linear Regression, Linear Regression based on selected variables, Lasso, Lasso on selected variables, Decision Tree, Bagging, Random Forest, Boosting based on their mean squared errors. Lasso on selected variables has the smallest mean square error:



By loading the 2020 loan applicants data and applying the best model on 2019 data set, I predicted the amount paid for each applicant in 2020 and calculated the profit by using predicted paid amount minus the loan amount and the predicted profits are in total 86072.3.