Loan Performance Analysis & Prediction

Juliette Zhu

Overview

The analysis leverages Application data, Loan Performance, US Zip Code data.

- Full scope of data covers 31 variables of 647 applications across 3 states during 6 month period.
- Loan Performance data includes first and second loan performances of applicants. This analysis only takes the first performance.

We segment all variables into 4 groups: Credit Score, Financial Status, Demographics, Application Details.

- Bad loan applicants are generally young, living a less stable life, having low credit score and monthly income, applying for larger and longer loans and splitting the payments into more numbers.
- Seasonality and regionality are perceived in bad loans.

Random Forest model is built to predict bad loan performance.

- Model has a 66% overall accuracy; predicts 88% of bad loans and 38% of good loans.
- Credit scores, Application date, Financial status, Age are the most important factors when predicting a bad loan.

Additional data points and dimensions will be helpful to the analysis.

- Unbalanced splits appear in the dataset. Larger amount of data is required to capture more variance of the true population.
- Other dimensions such as education level, household size, will be helpful to reduce model bias.

Table of contents

Bad loan applicants overview

Random Forest prediction model

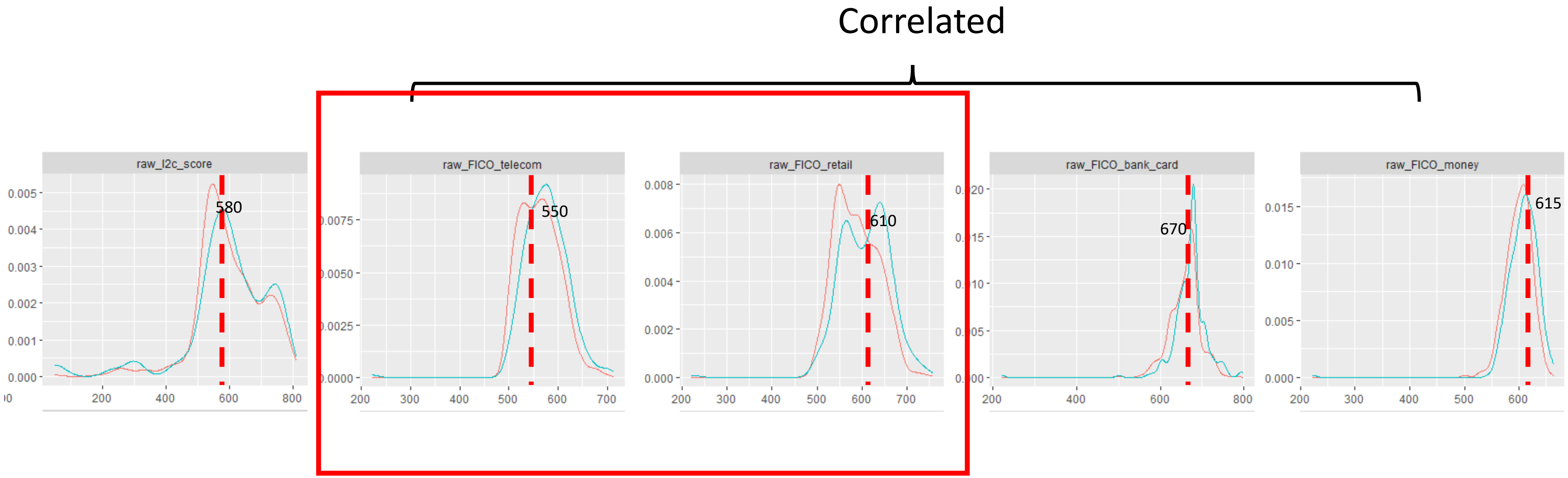
Data & analysis limitations

APPENDIX

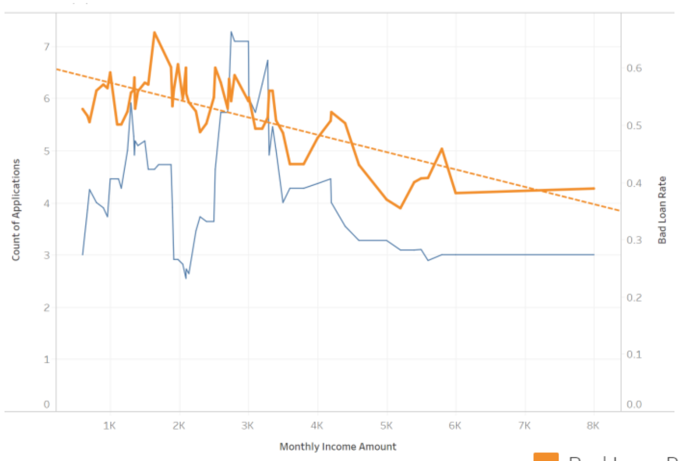
Bad loan applicants overview

Applicants with high credit scores and income are less likely to have a bad loan performance

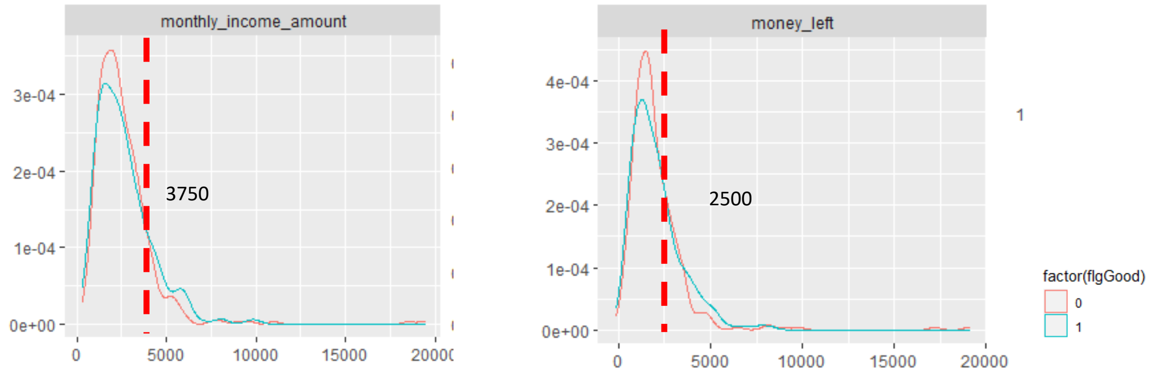
- Applicants with higher credit scores are less risky.
- Raw FICO scores across four categories are correlated with each other.
- FICO Telecom & FICO Retail scores are more differentiated between Bad/Good loan applicants.



- Higher Income group generally have lower Bad Loan Rate.
- Applicants having less than $2,500 after paying rent are more risky.



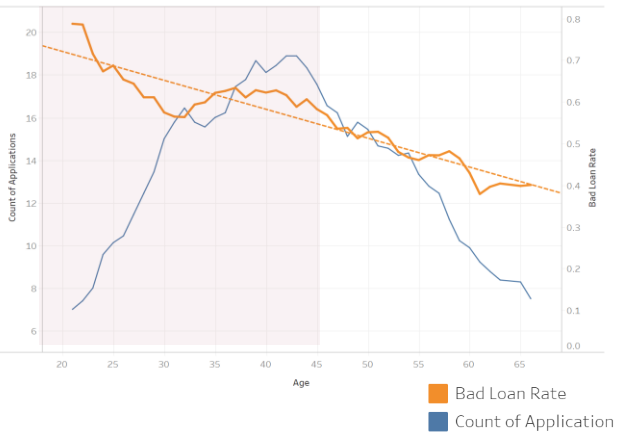
🟧 Bad Loan Rate
🟦 Count of Application

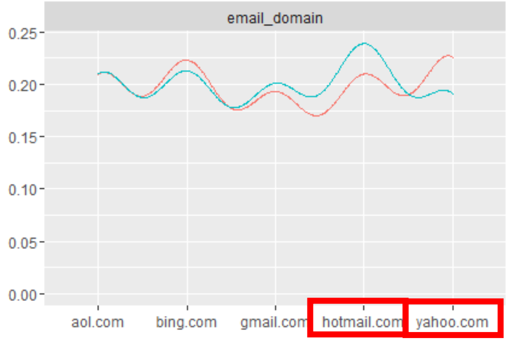


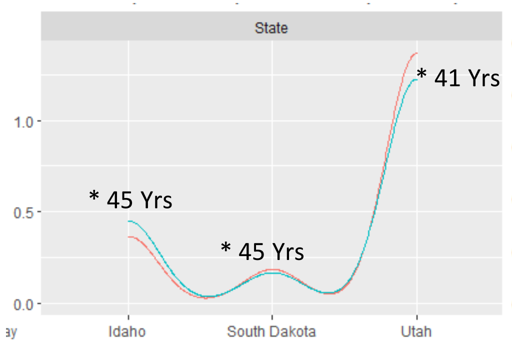
* Money_left = monthly_income_amount – monthly_rent_amount
* Correlation between raw_FICO_telecom, raw_FICO_retail, raw_FICO_bank_card, raw_FICO_money see appendix.
* Bad loan rate in Income graph is calculating moving average of +- 3.

Elder applicants with higher stability are less likely to make a bad loan; Regionality & email usage difference is perceived

- The risk of bad loans decrease with age.
- Hotmail users are more credible than Yahoo users.
- Lower bad loan rate is perceived in Idaho than Utah (probably due to higher median age).

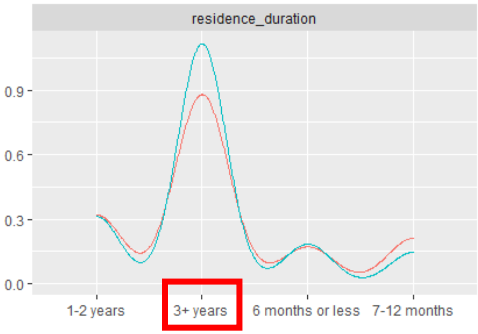


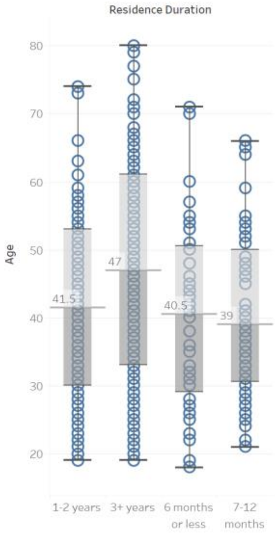


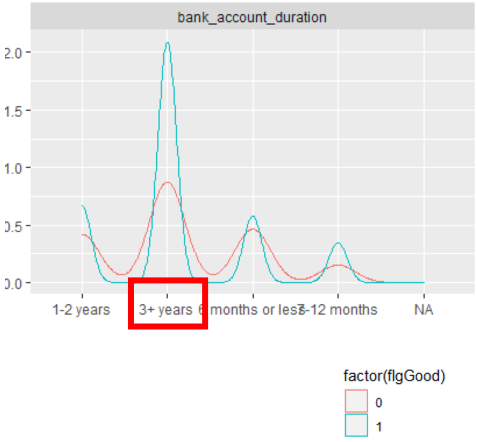


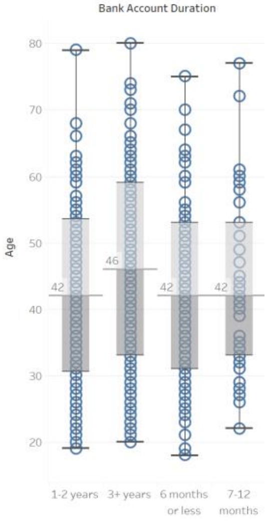
* State Median Age

- Applicants who rent/own current residence and have stayed for 3+ years are more credible.
- Applicants who have used bank account for 3+ years are much more credible.
- The duration could be related to Age.





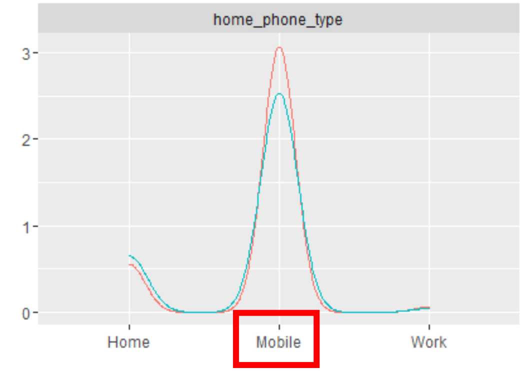


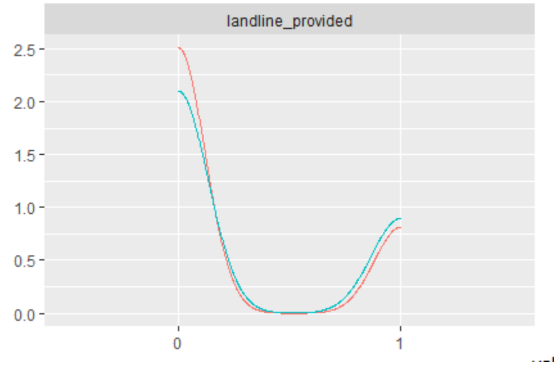


* Bad loan rate in Age graph is calculating moving average of +- 3.

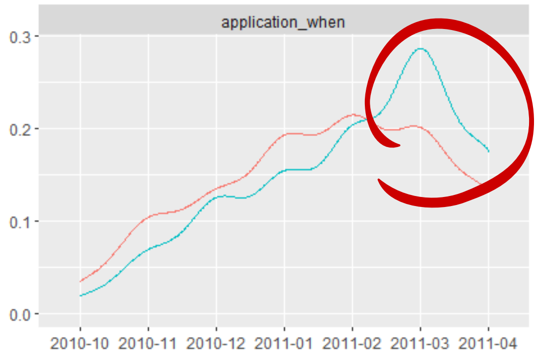
Applicants providing landline contact & depositing to bank account are less risky; seasonality is perceived

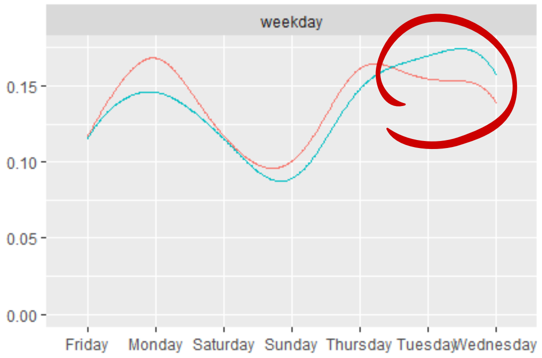
- Applicants providing landline contact are less risky.
- Providing Mobile number as primary contact is of higher risk than Home & Work.



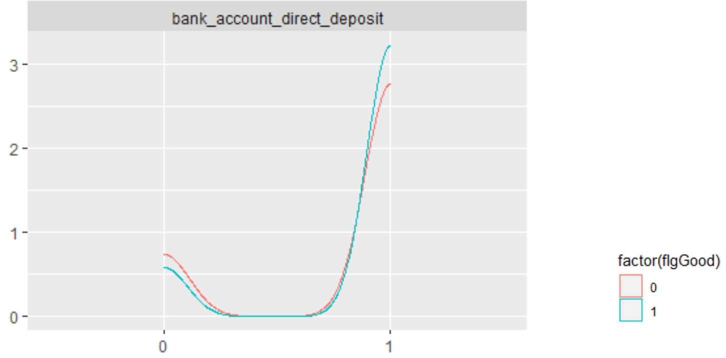


- Applications happen in March, April 2011 have a lower bad loan rate.
- Majority of applications happen on weekdays. The lowest bad loan rate is perceived on Tuesday & Wednesday, while Monday the highest.





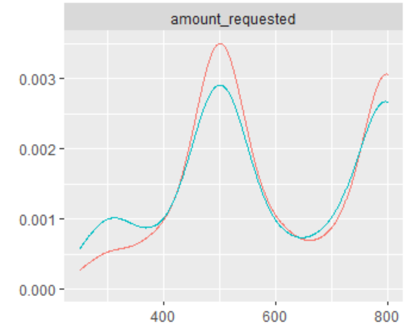
- Applicants deposit directly to bank account are less risky.

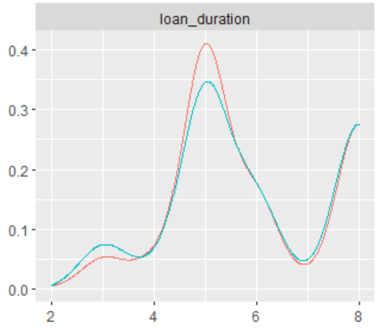


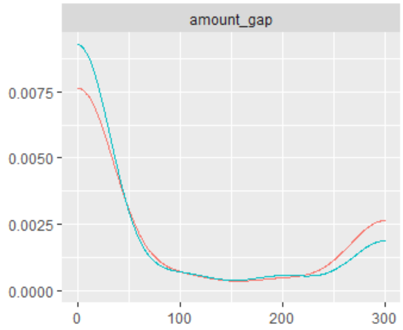
Small amount, short duration while less frequent payments of loans are less risky

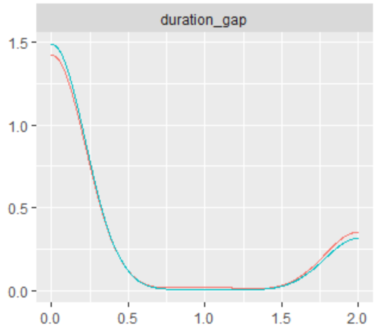
- Applications for small amount (< $400) & short duration (< 4 months) are less risky.
- Applications pre-screened as high risky (having a gap between requested amount/duration and approved amount/duration) are more likely to commit a bad loan.

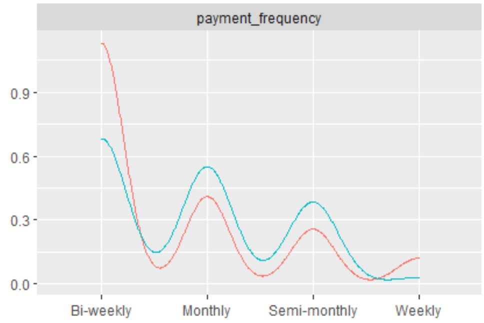
- Loans with more frequent & smaller amount of payments are more risky.
- Bad loan rate is generally lower at higher interest% of loan (related to payment frequency).

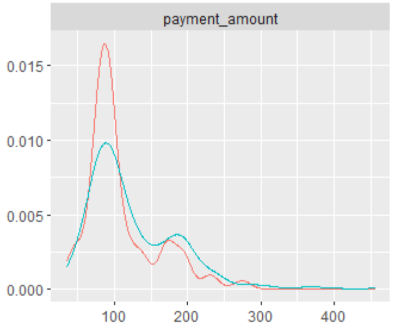


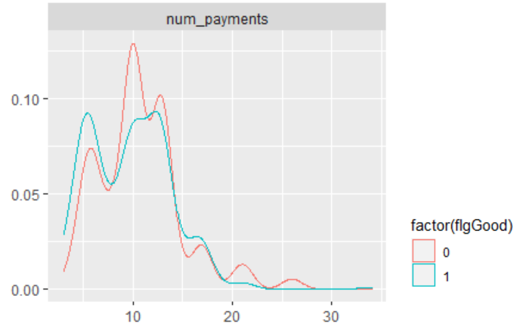


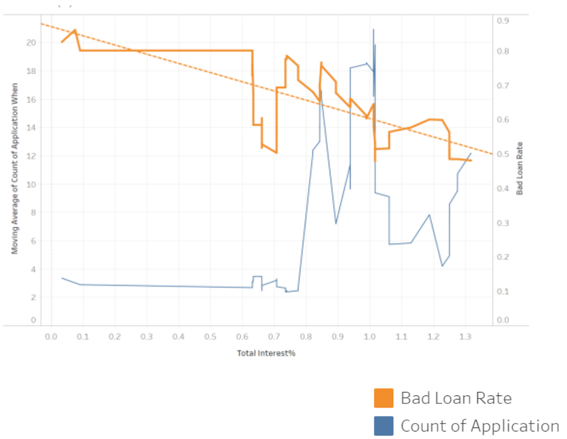










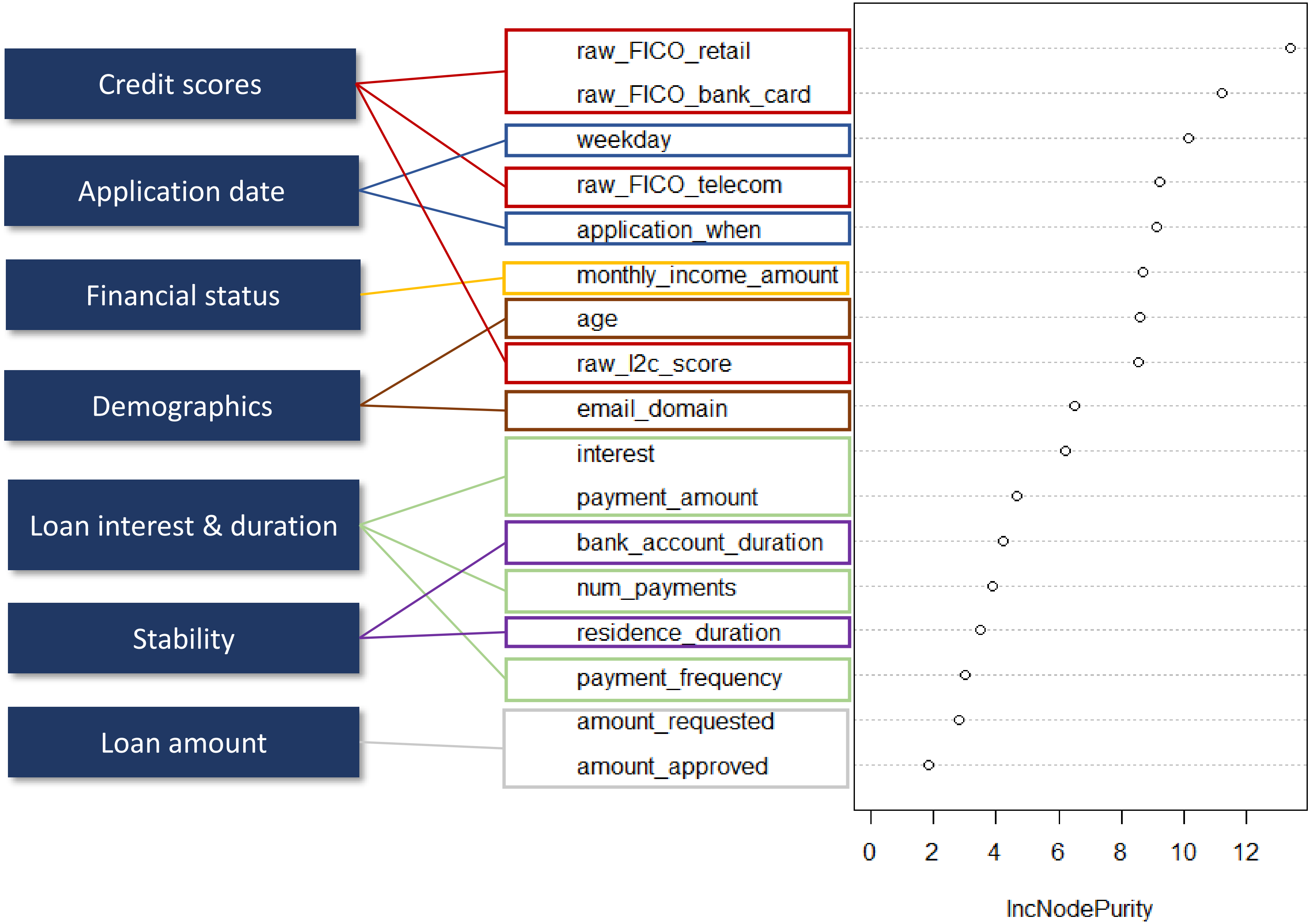


* Amount_gap = amount_requested – amount_approved, duration_gap = loan_duration – duration_approved
* Total Interest% = (payment_amount * num_payments – amount_requested) / amount_requested

Random Forest prediction model

Credit scores, Application date, Financial status, Age are the most important factors when predicting a bad loan

Random Forest model is built on variables that appear to have significant differences in the bad and good populations



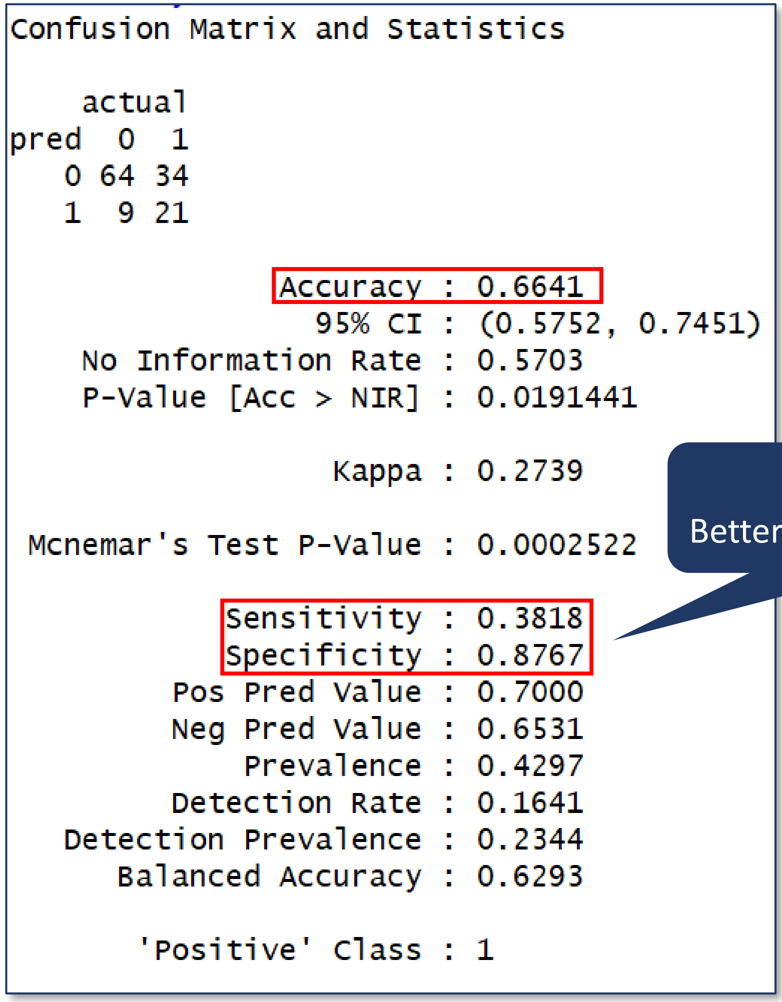
Random forest model has a 66% overall accuracy; prediction is false negative

Model Advantages

- Predict 88% of bad loans
- Work with both numeric and categorical variables
- Handle outliers & NAs properly
- Ensemble learning, more accurate

Model Limitations

- Require large sample size
- False negative, too strict for good loans
- Correlated variables will dilute importance



```
Confusion Matrix and Statistics

    actual
pred  0  1
   0 64 34
   1  9 21

               Accuracy : 0.6641
                 95% CI : (0.5752, 0.7451)
    No Information Rate : 0.5703
    P-Value [Acc > NIR] : 0.0191441

                  Kappa : 0.2739

 Mcnemar's Test P-Value : 0.0002522

            Sensitivity : 0.3818
            Specificity : 0.8767
         Pos Pred Value : 0.7000
         Neg Pred Value : 0.6531
             Prevalence : 0.4297
         Detection Rate : 0.1641
   Detection Prevalence : 0.2344
      Balanced Accuracy : 0.6293

       'Positive' Class : 1
```

False Negative
Better at predicting bad loans

Data & analysis limitations

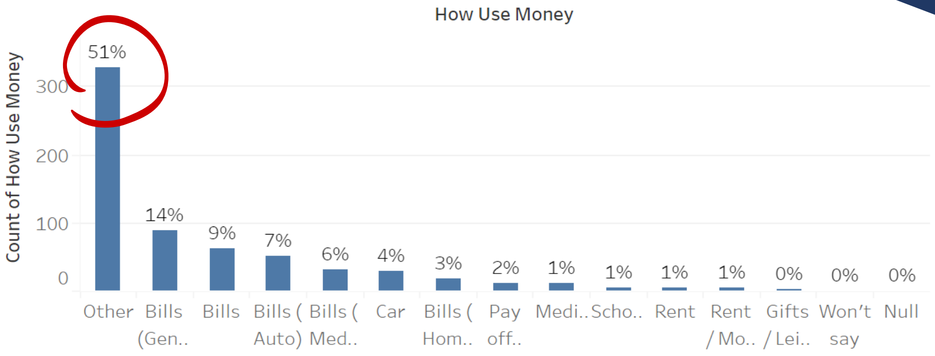
To improve prediction accuracy, we need more data points and dimensions

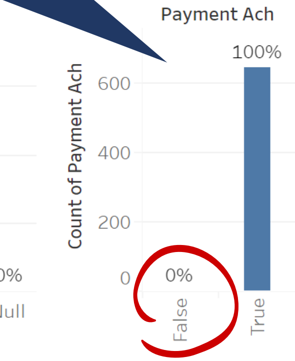
- Seasonality and regionality is perceived in the dataset. Need longer period and more states of data to verify the impact.
- Multiple splits in the dataset are unbalanced, which leads to model inaccuracy.
- Correlations and interactions are perceived. More data points are required to identify whether there is true effect or it’s due to sampling bias.
- More dimensions about the applications and applicants, such as education level, employment type etc., will be helpful. We can also integrate with other data sources such as US census data for regional financial status, bank account data for bank name and location from routing numbers.

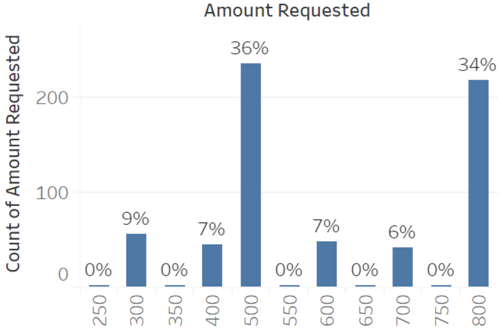
Only 1 out of total applications is not ACH Payment

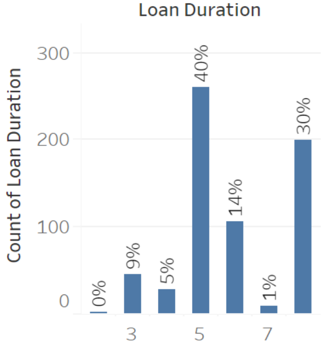
Monthly Payment applicants are avg. 8 years older than other groups

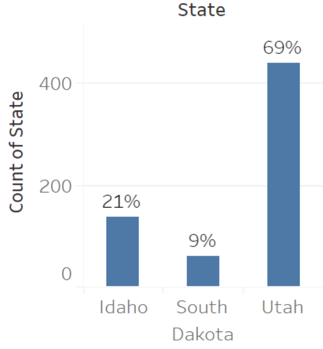
Utah state has more large amount of loans

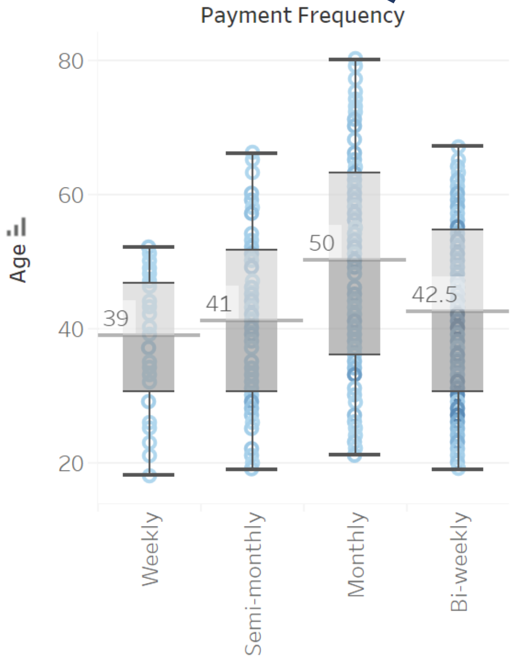


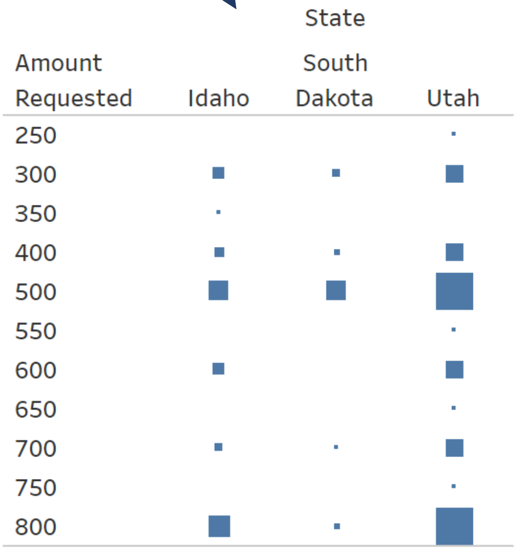






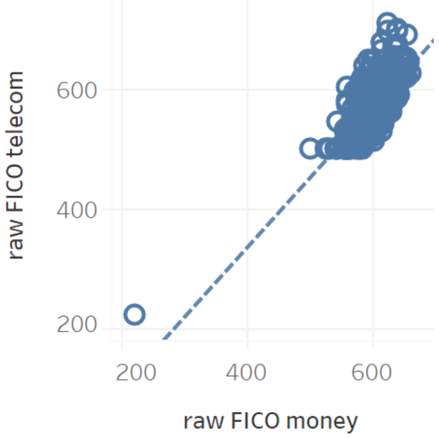


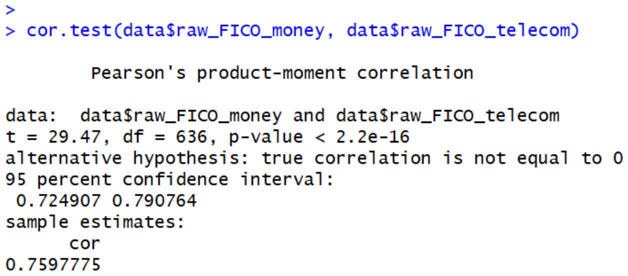




Appendix

FICO scores are correlated to each other, while no correlation is perceived between FICO scores and l2c score

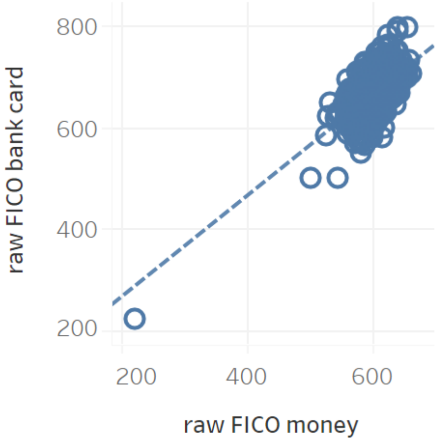


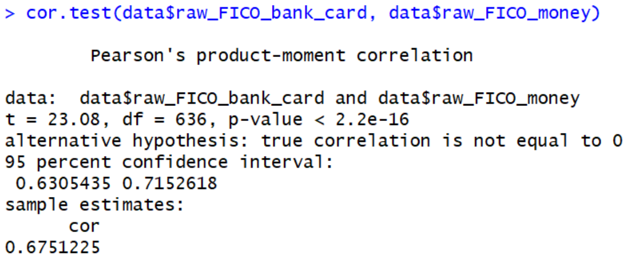


```
>
> cor.test(data$raw_FICO_money, data$raw_FICO_telecom)

        Pearson's product-moment correlation

data:  data$raw_FICO_money and data$raw_FICO_telecom
t = 29.47, df = 636, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.724907 0.790764
sample estimates:
      cor
0.7597775
```

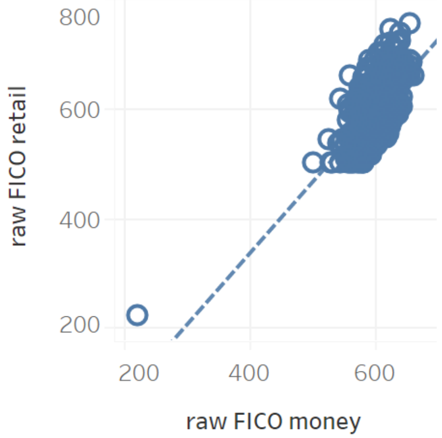


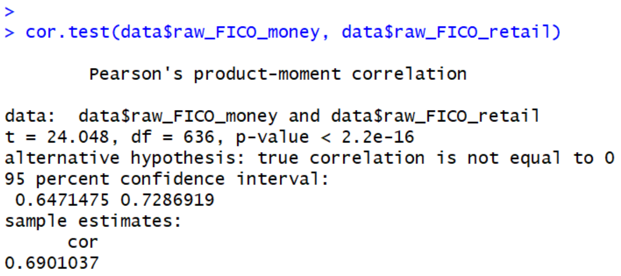


```
> cor.test(data$raw_FICO_bank_card, data$raw_FICO_money)

        Pearson's product-moment correlation

data:  data$raw_FICO_bank_card and data$raw_FICO_money
t = 23.08, df = 636, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6305435 0.7152618
sample estimates:
      cor
0.6751225
```

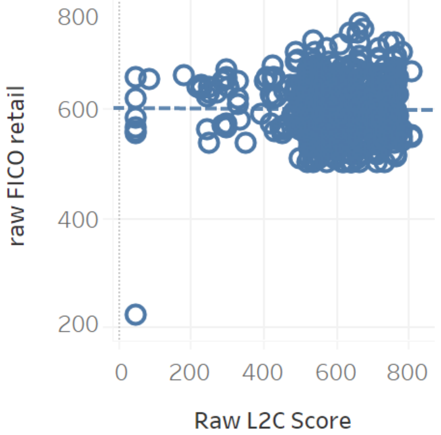


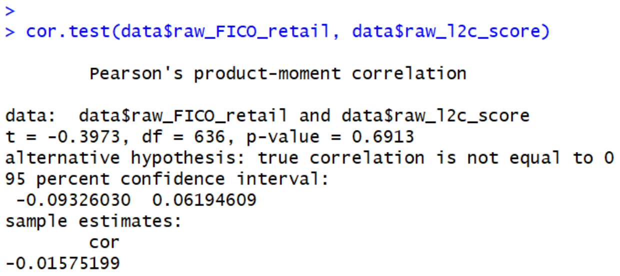


```
>
> cor.test(data$raw_FICO_money, data$raw_FICO_retail)

        Pearson's product-moment correlation

data:  data$raw_FICO_money and data$raw_FICO_retail
t = 24.048, df = 636, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6471475 0.7286919
sample estimates:
      cor
0.6901037
```





```
>
> cor.test(data$raw_FICO_retail, data$raw_l2c_score)

        Pearson's product-moment correlation

data:  data$raw_FICO_retail and data$raw_l2c_score
t = -0.3973, df = 636, p-value = 0.6913
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09326030  0.06194609
sample estimates:
        cor
-0.01575199
```