Assignment A3.1 - Analyzing a Sharing Economy Dataset

Analysis developed by César Alejandro Martínez Ortíz

Analysis preview

Dataset summary description: "Historical loans from both Prosper and Lending Club from 2013 - 2018."

Pipeline analysis summary:

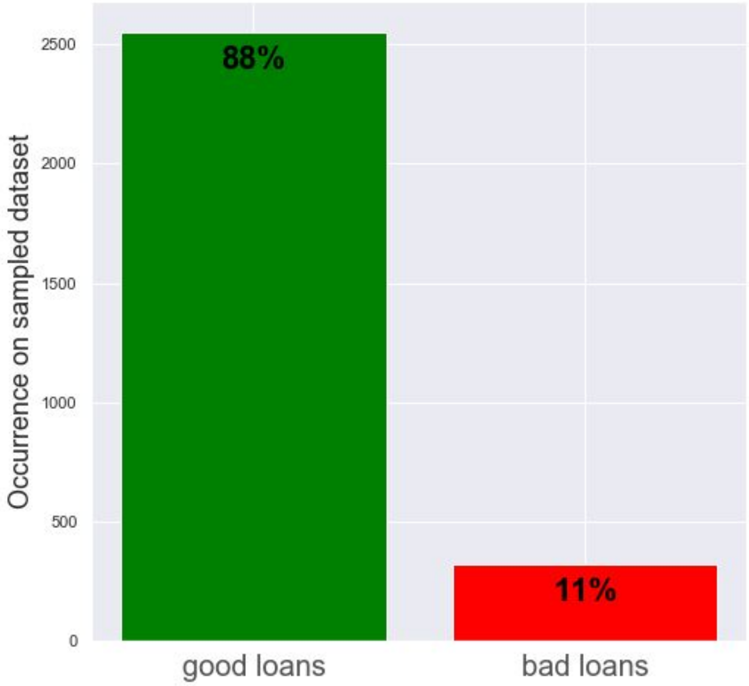
- Features were reduced first by intuition and then by analysis to reduce noise and provide business-actionable classification insights.
- The 'Loan status' description was simplified to consider **'Completed'** or **'Current'** loans as **'Good for business loans'** and all others as '**Bad for business loans**' (**Defaulted, Charged Off** or **Cancelled**)

Research question: ***¿How to tell apart good loans from bad loans?***

***I took the liberty of not only separating the defaulted loans because charged off and canceled loans are also bad for business and assumed they could be grouped. The summarized, but more detailed analysis pipeline can be found as an appendix at the end of this presentation.**

Insights from exploration

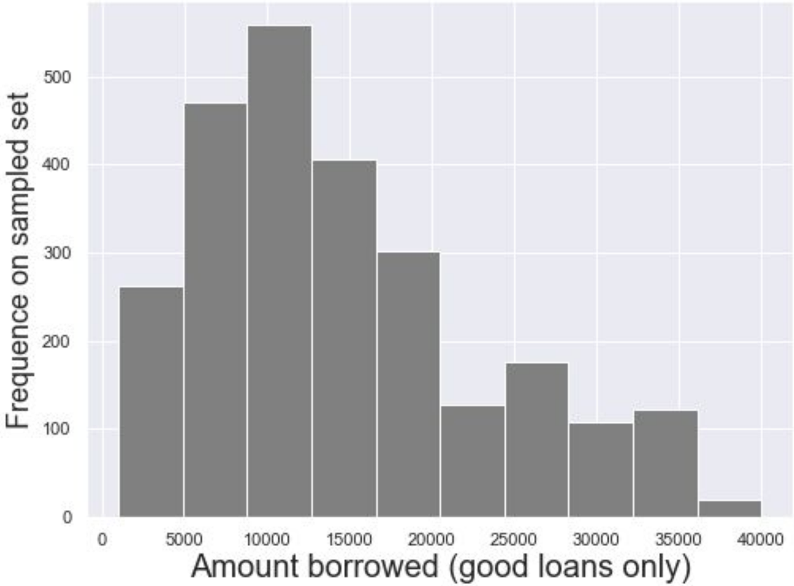
Most loans given are ‘good loans’, which means that in practice, the process of loan granting selection has a high chance of giving out loans that will be paid successfully.



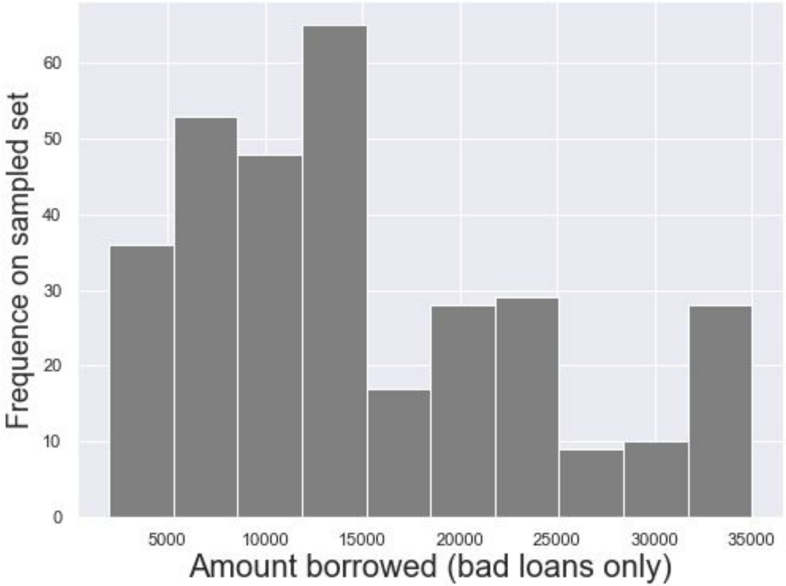
Insights from exploration

Nevertheless, **'Bad loans' still happen and they still mean money**, efforts should be made to minimize this risk as much as we can as they still are meaningful for business.

Sum total of 1% sample: $ 37, 404, 531

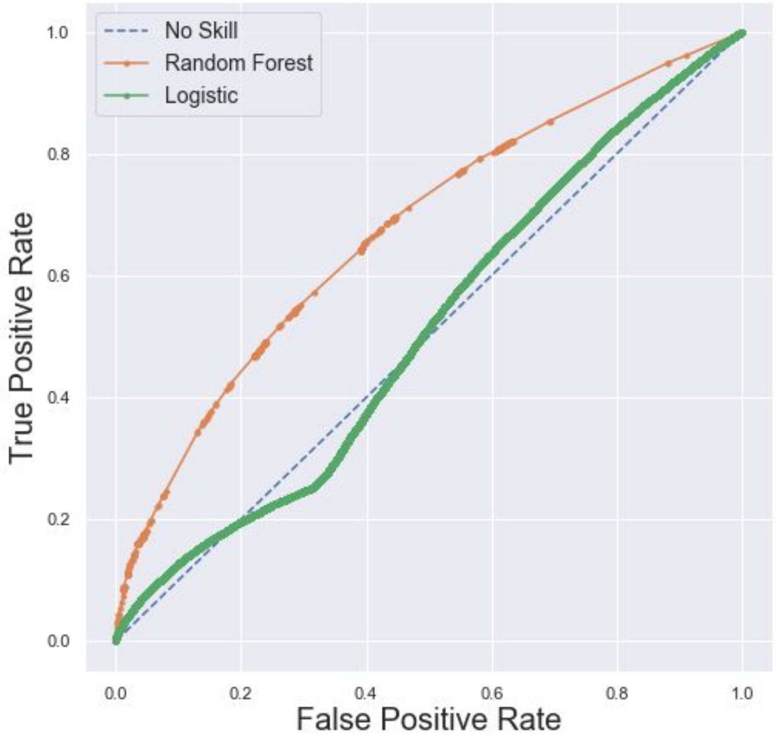


Sum total of 1% sample: $ 4,941,460



Model selection summary

To achieve this classification task, random forest trees and logistic regression were tested against a 'No-Skill' classifier. Random forest was selected as the one that had the highest ROC AUC score.



ROC-AUC scores

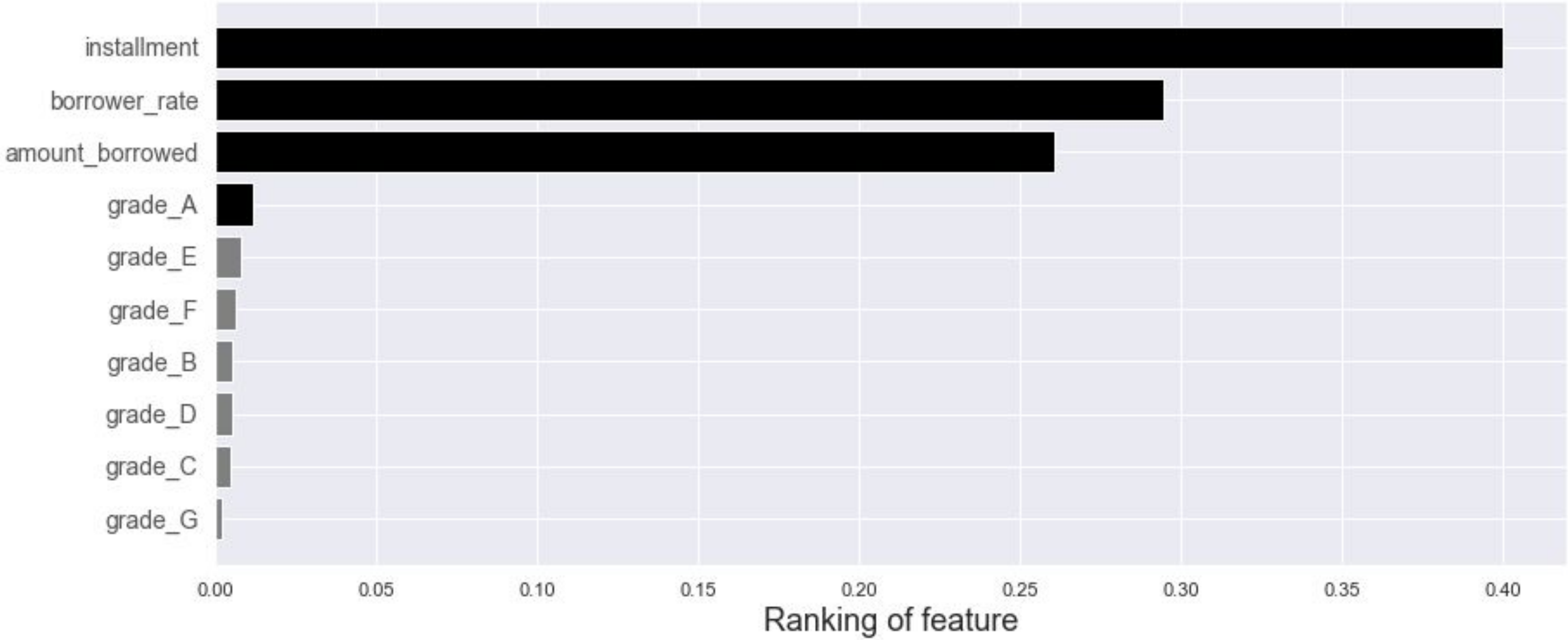
No Skill (reference): 0.500

Random Forest: 0.675

Logistic: 0.511

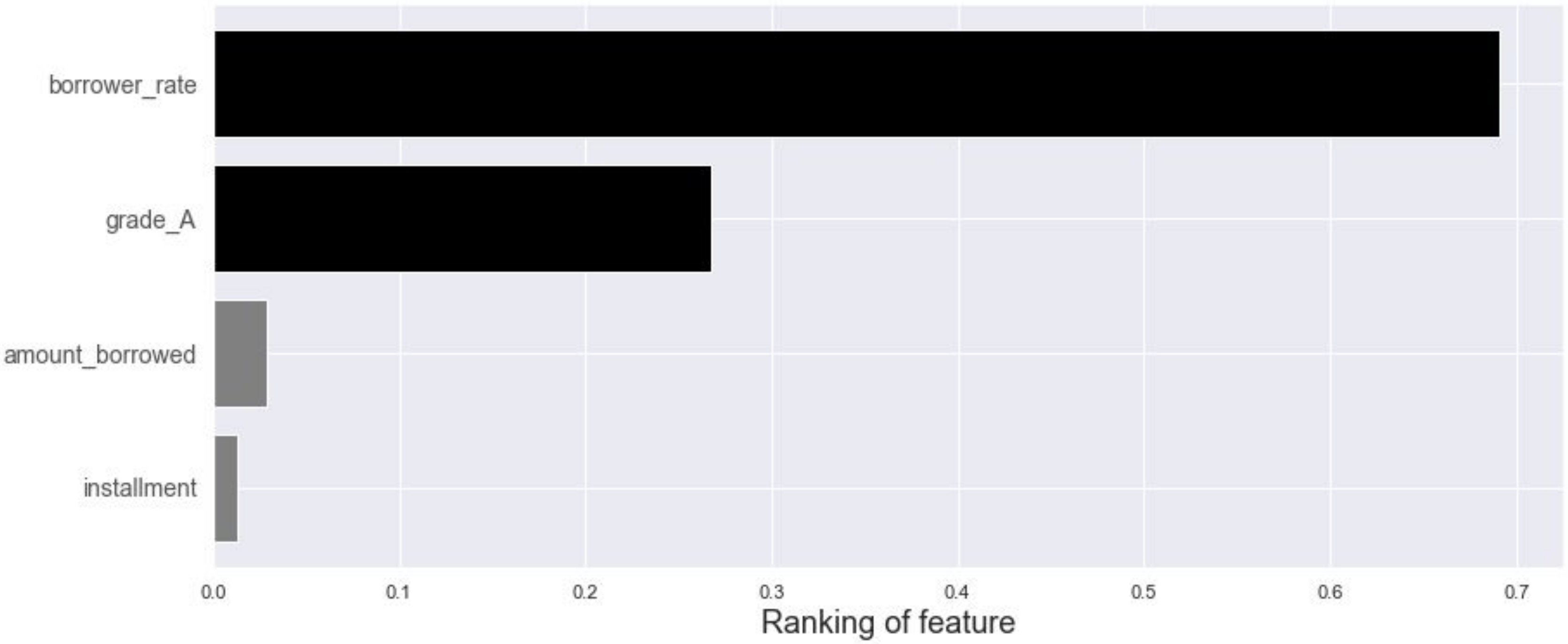
*Further tuning could be made but this was left out of the scope for this analysis.

Insights gained from model (pre-grade reduction)



- Risk Credit Grades below A have very similar importance (we can simplify this separation task as having a Risk Credit Grade of A or not A)
- Installment, Borrower rate and Amount Borrowed have very high impact on classification prediction for this model

Insights gained from model (after grade reduction)



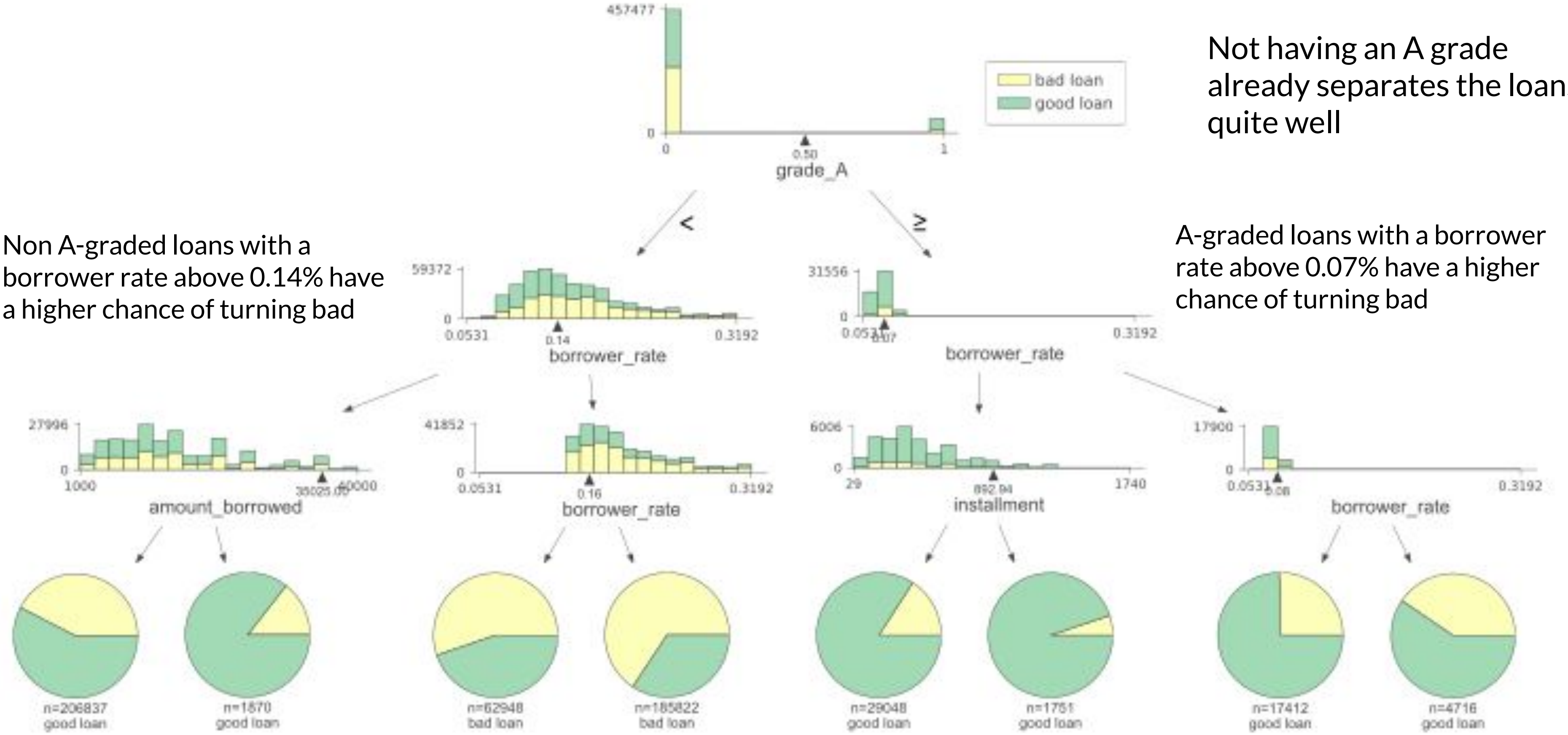
Amount borrowed and installment dropped in feature importance after grade reduction, this is an interesting insight. Grade alone and Borrower rate could be used to pre-screen loan predicted fate in a reduced-feature model.

Decision tree visualization (Extended insights after)

Not having an A grade already separates the loans quite well

Non A-graded loans with a borrower rate above 0.14% have a higher chance of turning bad

A-graded loans with a borrower rate above 0.07% have a higher chance of turning bad



Insights gained from decision tree

An A-graded loan predicted to be a bad loan has a high chance of being a good loan in practice. Such cases should be reviewed by an expert.

Loans graded below ‘A’ have higher chances of being bad, this could be treated as a good indicator for fast pre-screening.

Installment, borrower rate and amount to be borrowed for loans graded below A can act as additional decision boundaries for refined-screening or as cautionary recommendations to provide loans with lesser possible impact (i.e. lower borrowed amount, higher installment, finely tuned borrower rate)

Loans graded above A with a borrower rate below 0.07% have a very high chance of be completed and installment and amount to be borrowed do not have too much weight on the outcome in this case.

Appendix