Moral Loan Delinquency and Default Prediction

Yara Mubarak, Yang Shi

April 2019

Other Group Members: Jordan Fan, Hsiao-Yu Chiang, Florin Goodluck Langer, Bizuayehu Whitney

1 Introduction

With the introduction of new data sources and computing power, the issue of morality presents itself in many data science fields. We aim to test whether ”moral” models can perform as well as models that include all kinds of data, including data that might be a result of immoral social policy. Furthermore, historically many communities have been marginalized, causing incomes to de- crease, debt to increase and by consequence the ability to take loans and pay them off decreases. These communities share common indicators of area, low income, and high debt. While in a completely fair world, these indicators would be void of sociopolitical status, in our world, they aren't. Therefore, in this project, we aim to test whether we can build high-quality models that don't take into account these sociopolitical factors and rather focus on the individual's past delinquencies and the type of loans they have been taking out. While we cannot perfectly design models that do not take into account sociopolitical factors, we can design them to focus more on the individual in a community rather than the community itself. That would aid in removing unjust sociopolitical factors in lending decisions.

In this project, we are working with the Lending Club's Kaggle Dataset. The Lending Club is a company that allows peer-to-peer lending, where borrowers can apply for personal loans, auto refinancing, business loans, and elective med- ical procedures without going through big banks and red tape. They have made their data available for the public and that is what we are using. Through this data, we aim to design classification models that allow us to predict whether a user will be late in their loan repayment or completely default versus being able to completely pay off a loan without issues.

By using ensemble learning like boosted trees, bagged neural nets and simple linear regression, we focus on examining which model complexity will align with the data. We will be testing two kinds of models: One model that includes all the indicators and another which we strip of demographic data like income, area code, and past debt.

2 Data Cleaning and Preprocessing

We define an instance of our data as one a loan being taken out. Our features included 145 columns about the loan and the user taking it out. Some of the features were devoid of data due to privacy issues, like user ID and transaction ID. We ~~of course~~ removed those and went on to work with our data. In the end, we ended up working with 75 features, some of which were one-hot coded. These will be discussed in the section below.

2.1 Creating Data Labels

Since our goal was to predict whether a certain instance was ever paid late or defaulted, we had to use a combination of original features from our dataset to do that. We used loan status, which had the following entries about the current status of the loan, with ones meaning they count as late or default and 0s being they are currently paying without hitch:

- Charged Off (1)
- Current (0)
- Default (1)
- Does not meet the credit policy. Status: Charged Off (1)
- Does not meet the credit policy. Status :Fully Paid (0)
- Fully Paid (0)
- In Grace Period (1)
- Late (16-30 days) (1)
- Late (31-120 days) (1)

Then we used past indicators of being late using the following features :

- Delinquency amount
- Hardship Reason

A hardship is taken out on a loan because of difficulties paying the loan, which means that it was late in the past and the user had to petition to be able to pay it late. ~~A~~ delinquency means that the user had not paid off the loan before and did not petition for hardship. Furthermore, if any of these columns had non-zero entries, we would use that to label that instance as a loan that had been late in the past.

2.2 Cleaning the Data

Many of the 145 features had to be written off as they had been descriptive of hardships and delinquencies, which we wanted to use as labels. Therefore they could not be used as features. Some of the features were also exclusively used for data exploration because they could not be feasibly integrated into a numerical model. We had two models, as discussed in the introduction. One where we used all the data and the other where we focused on using moral data.

Below are the loan features we used :

Loan Data Features		
Feature Name	Feature Type	Feature Description
Funded Amount	float	How much money was requested by the user?
Invested Amount	float	How much money was committed by the investor?
Installment	float	The periodic installments of the loan.
Interest Rate	float	The interest rate on the loan.
Total Payment	float	The total current payment of the loan.
Issue Data	float	Date of issuance by days from $1^{s}t$ of December (the latest date in the dataset).
Payment Plan	One-Hot	If the loan was to be paid off with installments or all at once.
Loan Term	One-hot	Whether the loan was 36 or 6 months.
Grade and Subgrade	One-Hot	One hot indicator of the type loan grade according to the Lending Club.
Loan Amounts	float	This is how much money in $ the loan was.

Table 1: Features that pertain to loans and their description.

Below are the user features we used and Model B is the moral model:

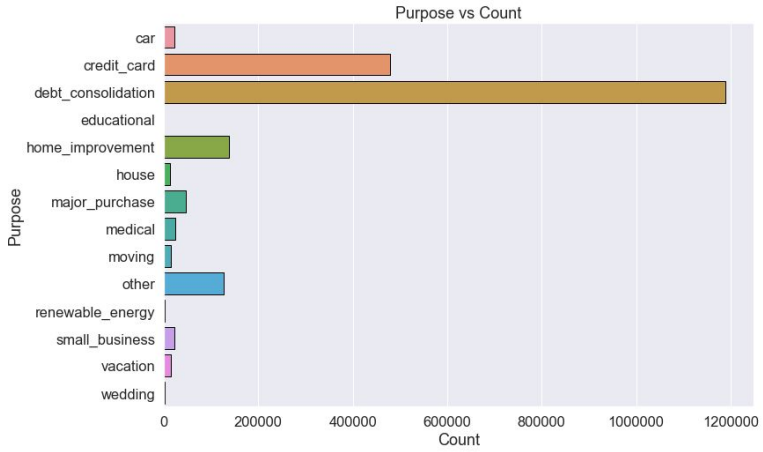
User Data Features				
Feature Name	Feature Type	Feature Description	Model A	Model B
Source Verified	One-hot	if the user was verified by the Lending Club or not	yes	yes
Job Title	string	Each user's job title.	no	no
Time at Job	integer	The amount in years the user has been at their job.	yes	yes
Home Ownership	One-Hot	Whether the user owns, rents, or has a mortgage on their home or none.	yes	no
Annual Income	string	The annual income of user taking the loan.	yes	no
High Credit Limit	string	The total high credit limit of the user	yes	yes
Chargeoff	One-hot	If the user has been charged off a loan in the past year.	yes	yes
Bankruptcies	integer	The number of public record bankruptcies the user has.	yes	yes
Earlierst Credit Line	float	Number of days from December 16*st* of the user's earliest credit line	yes	yes
DTI	float	Debt to Income Ratio of user	yes	no
Delinquincy	integer	The number of delinquencies the user has had in the past 2 years	yes	yes
Time to Last Delinquincy	int	Number of months since the last delqinquency of user	yes	yes
User State	string	User's state code.	no	no
Purpose	string	User's purpose of taking the loan	no	no
Zip Code	float	The first three numbers of the user's zipcode	yes	no

Table 2: Features that describe the lender and their description. Model A refers to the holistic model while model B refers to the moral model.

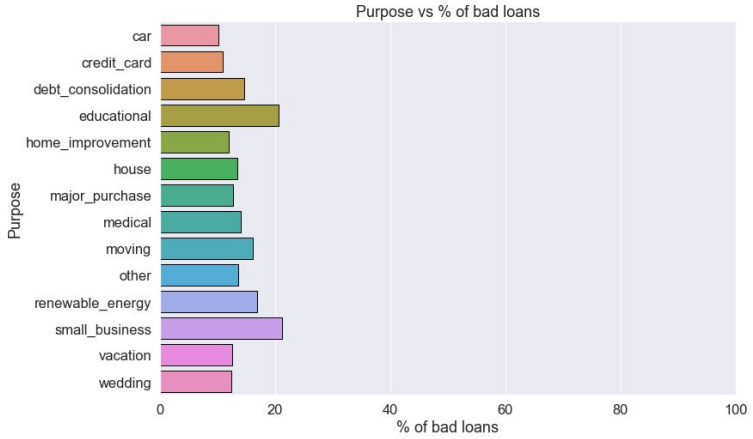
3 Data Exploration

3.1 Purpose Exploration

In this section, we explore why people have taken loans and how that affects how late they will be.



(a) Purpose Histogram



(b) Percentage Late Barplot

Figure 1: Histogram of the Purpose instances in the Lending Dataset

As seen in Figure 1 there seems to be an overwhelming number of instances of taking loans to consolidate other debt, with either credit card or other loans. However, by a margin of 5%, educational and small business loans overtake them with late loan percentage.

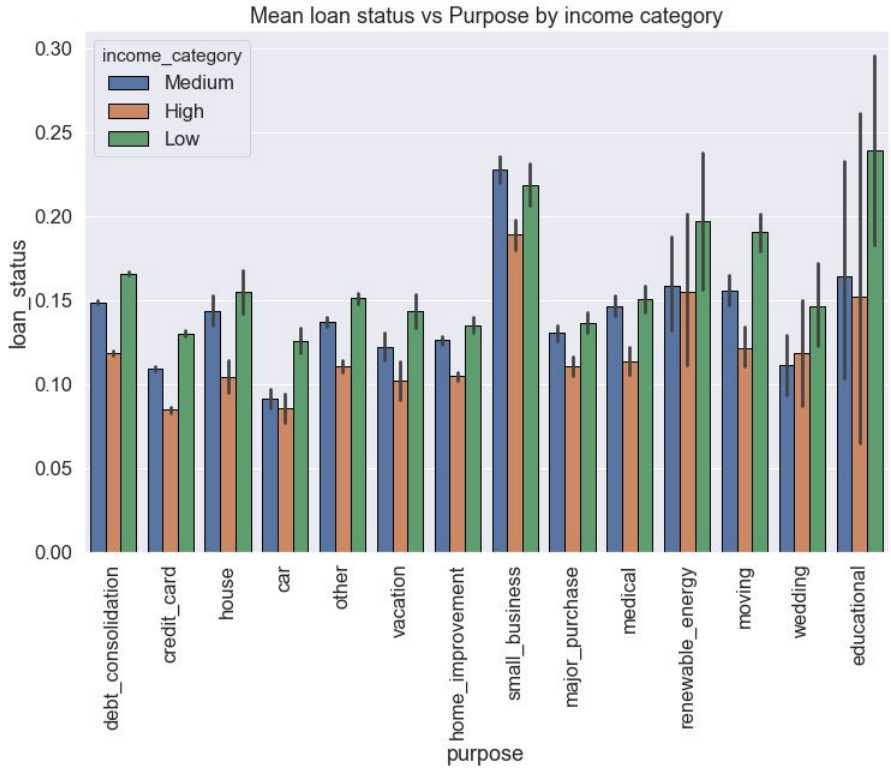


Figure 2: Late Payment by Purpose and Income Barplot. The y-axis shows the percentage of late payers. The axis is spread into purpose and then income subcategories.

As shown in Figure 2, the highest group of people not being back loans are low in- come categories taking loans for educational purposes. Averaging at about 22%, just below educational top 24% late payers, small business show no disparity between income levels but rather an equally high percentage of late payments for all.

3.2 **Income Exploration**

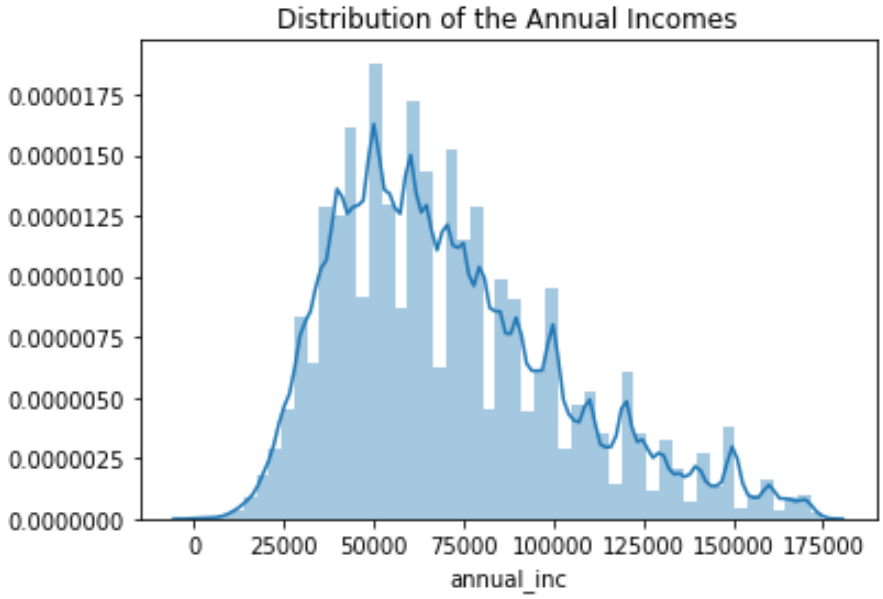


Figure 3: A histogram of annual income in $s

As seen in Figure 3, we have a left tailed distribution that peaks at around 50k annual income and goes all the way up to 175k. As expected, people with higher income take higher peer to peer loans.

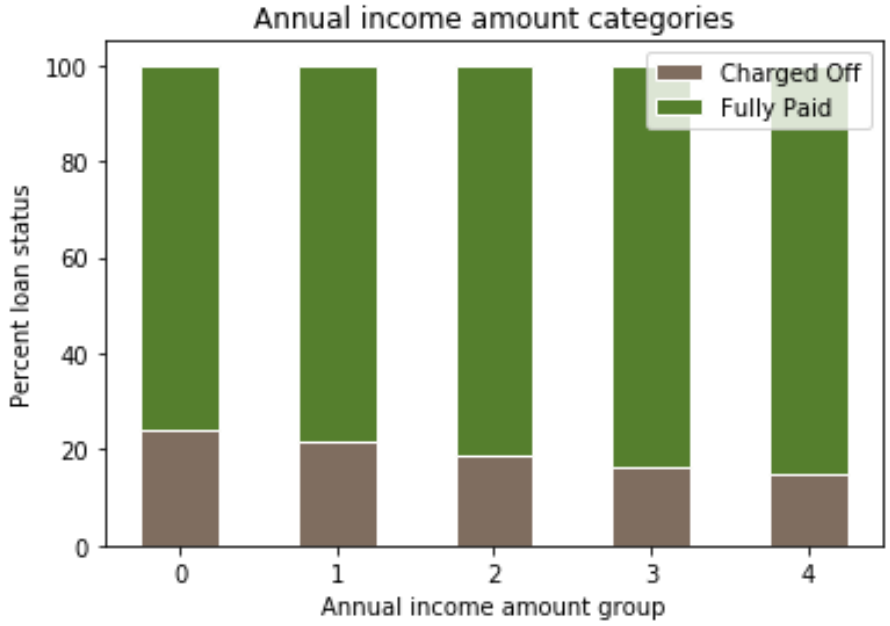
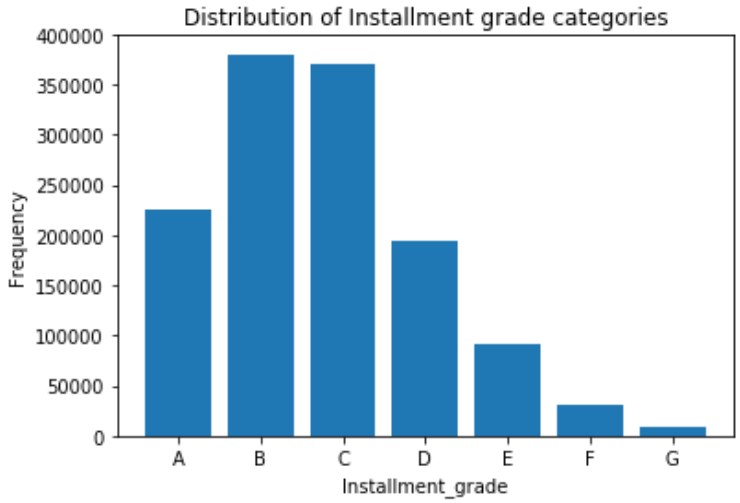


Figure 4: A histogram of annual Binned income with the percentage of late payers

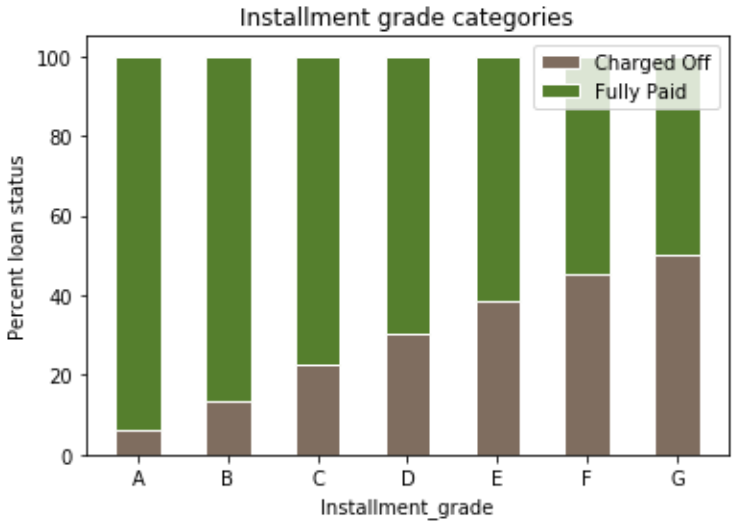
As shown in Figure 4, the percentage of late payers has a negative correlation with annual income as there is a clear downward trend in the percentage of as we move up the groups.

3.3 Loan Grade Exploration

As seen in Figure 5 We can see that there is a strong and clear trend between the percentage of late payers and the grades as we move from grade A to G. We expect both our prediction algorithms to pick up on this. We can also see that loan decision makers have realized that G grade loans are very risky and therefore give a low amount of them.



(a) Grade Histogram

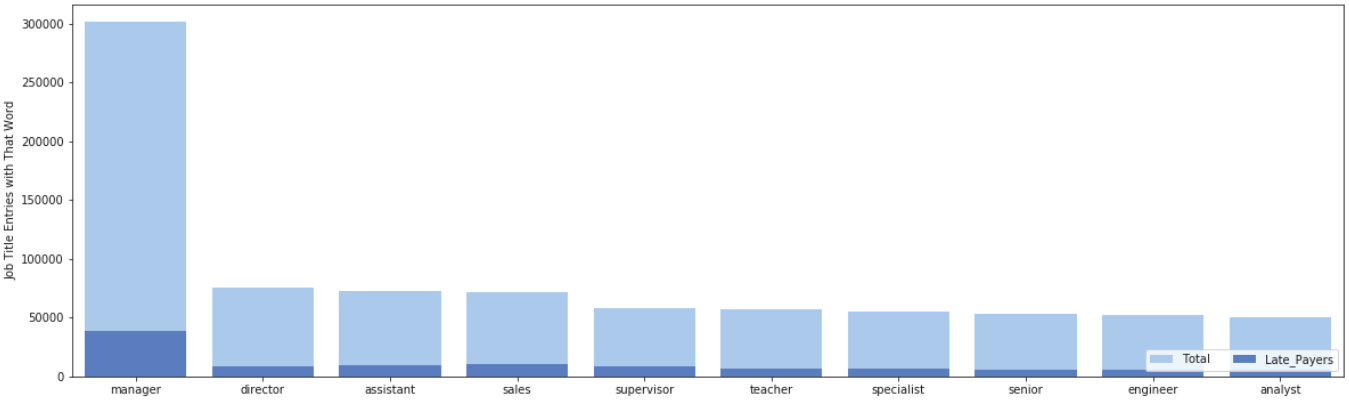


(b) Percentage Late Barplot

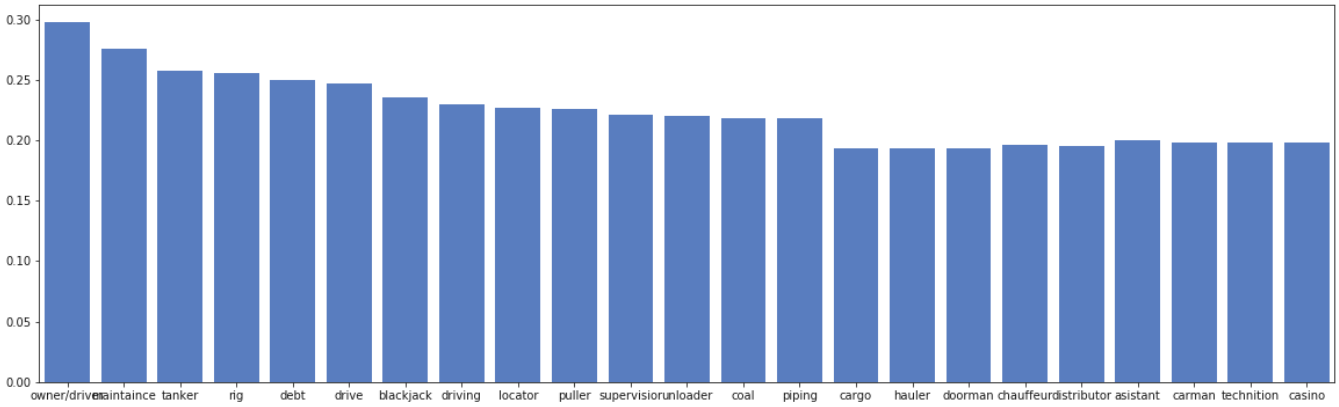
Figure 5: Histogram of the Grade instances in the Lending Dataset

3.4 Job Title Exploration

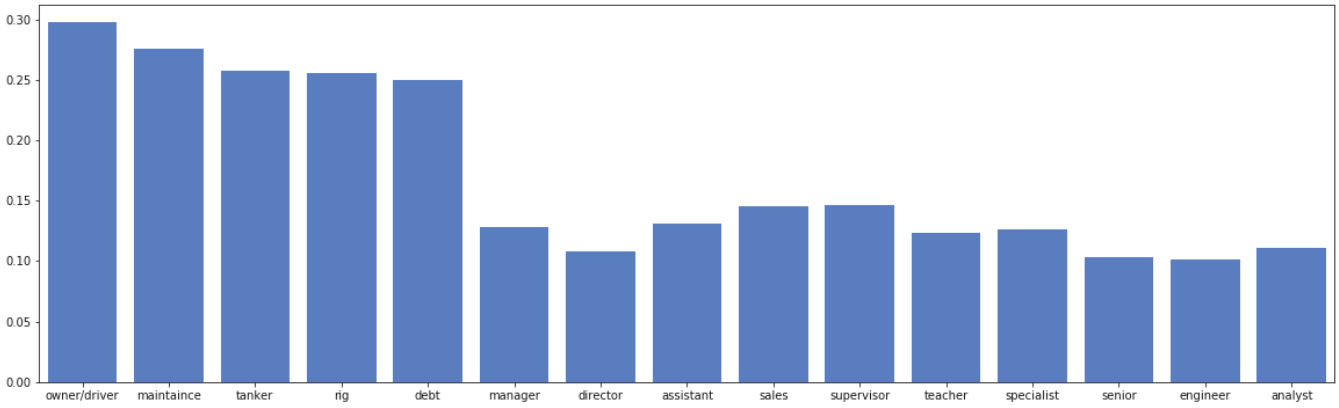
As seen in Figure 6 we have jobs who are on average twice more likely to not pay their loans on time. These job titles are in the fields of driving (as seen by the driver, driver/owner, trucker, etc...) and gambling (casino and blackjack).



(a) Top 10 most frequent words in Job titles and their late payers



(b) Titles With the Latest Payer Percentage



(c) ”Good” vs ”Bad” Job Titles

Figure 6: Exploring the words in job titles to see their frequency and the per- centage of their late payers.

4 Prediction Models Their Results

Following all the exploratory work done, we decided to experiment with models as well. Although there seems to be a strong linear trend between grade and bad loan takers, it is still not enough to have an accurate predictor. We tried the following models listed with increasing complexity below:

Models and their scores				
Model Name	Accuracy A	False Negative Percentage A	Accuracy B	False Negative Percentage B
Linear Regression	NA	87%	—	—
Random Forest	96 %	2.724 %	—	—
Boosted Gradient Classifiers	95.9 %	0.14 %	95.9%	0.14 %
Bagged Neural Nets	86.56%	0 − 0.5%	86.55%	0 − 0.5%
Adaboost Trees	96.5 %	0.46%	96.6 %	0.42 %

Table 3: Models and their Accuracies as well as False Negative Percentage. Model A refers to the model incorporating all the data while model B refers to the ”moral” model

Here we are concerned with the percentage of false positive because, with every misclassified false instance, the most cost is incurred. Moreover, a neg-ative represents that the borrower will not be late or default. The number of false negatives is the number of times the predictor said that a borrower won't be late, but the borrower ended up being late or defaulting. Logically, that is where the lender loses the most. In other cases of bad prediction, the only cost incurred can be being extra conservative. Relatively, the cost of being con-servative is very low compared to giving out loans that will end up in default or late payments. Keeping that in mind, we would like our predictors not to be too conservatives, because that will result in inefficient usage of the lending platform. Moreover, lenders cannot lend as much as they'd like and borrowers cannot borrow unless it's very low risk. That is why we keep accuracy as an overall metric as well.

We will delve deeper into the best model overall which seems to be the gradient boosted trees. With an accuracy of 96%, only 1% below that of the highest AdaBoost trees, it seems to be a very high accuracy. What makes it the best, is that it has a very low number of False negatives. Therefore, we will take more about this model.

However, before we explore deeper, the issue of validation needs to be addressed. Given 2 million entries in the dataset, we found that it is sufficiently large such that we could divide it into 70 % training and 30 % testing without loss of generality or overfitting. Moreover, in the following sections, we will show how closely the training error follows the testing error. That is because we are training our algorithms on a very diverse dataset in terms of the data patterns. Therefore we found it unnecessary to use

validation techniques such as K-Fold or leave one out validation.

4.1 Boosted Gradient Trees

As seen in Table 3, Boosted Gradient Classifiers have a relatively high accuracy of 94.8% but by far the lowest false negative rates. We used 100 classifiers and a max depth of 1 for each decision tree. These classifiers are very simple, as is the case with the many boosting algorithms. Therefore with each gradient boost, they are able to fix for the error for hard instances but not overfit. Here, we define a hard instance as something that deviates from the normal pattern.

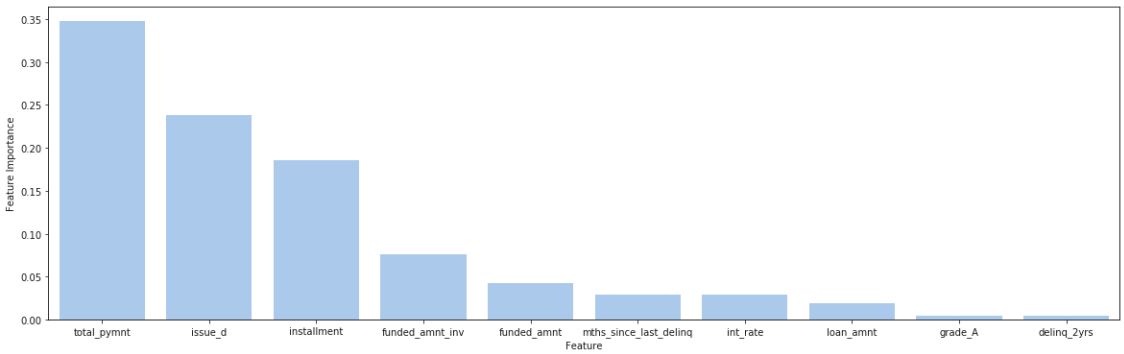


Figure 7: Top 10 features and their weighted importance according to the Gradient Boosted Trees in the holistic model

As shown in Figure 7, the most important features that seem to be the most important are the loan features as opposed to the user features. That is what accounts to the very low discrepancy between the moral model and the holistic model. In addition, we can see that the highest feature importance is the loan amount. This says that the loan amount will play the biggest factor in whether a lender will repay the loan on time or not.

We also investigated how our accuracy is increasing with each boost. Plotted in Figure 8, it can be seen that the testing along with the training accuracy plateau at around 96 %. Moreover, training and the testing accuracy seem to be very closely tracked, sometimes overlapping. That means that our training set is very diverse and that due to that, our model is not able to overfit.

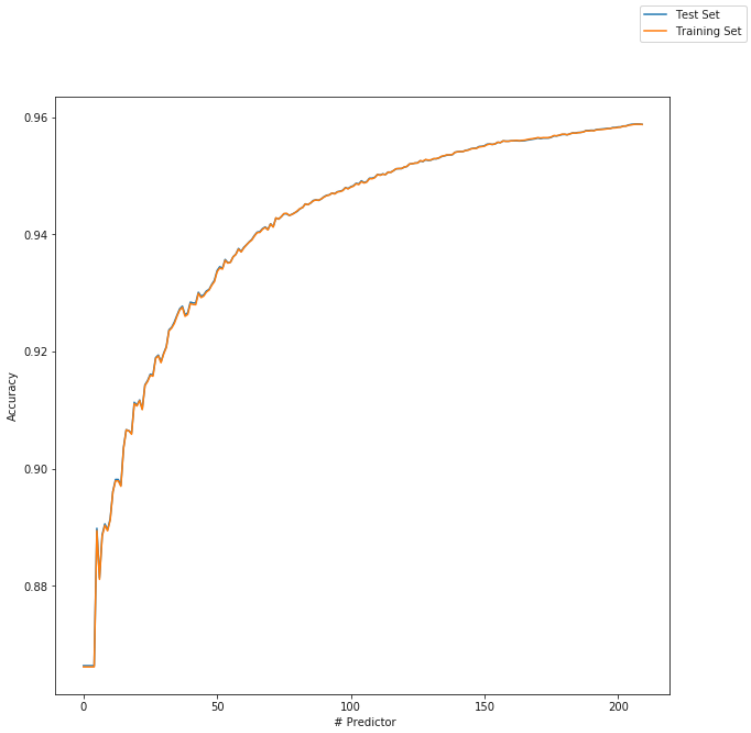


Figure 8: The Training and Test accuracy as we increase the number of boosted gradients

5 Conclusion

In conclusion, we found that the best model was a set of 150 gradients boosted trees. With 96% accuracy, we can predict whether a lender will default or have late payments. We also can say with certainty that we have less than 0.2 % False negatives. That means that 99.8 % of the time if our model says that the borrower will not default, they will not. Additionally, our moral model performed equally well, if not better, in some instances, than the holistic model which included the demographic data. Now, tying together our data exploration and our model we can come to the following conclusions:

- Loans for small business and educational purposes have the highest late and default borrower percentage. Especially educational loans for low- income borrowers. However, due to a mix of other features, our trained model did not reflect that. That is good news because, before the model, this finding could probably influence lenders and decision makers to stop lending to low income students or startup founders. That would not be beneficial to society. Therefore, our moral model would give an alternate route. The alternate route would instead affect the type of loan they are giving, rather than to whom they are giving it.
- Income always seems to be a daunting factor for lenders and borrowers. However, our trained model, as opposed to the downward trend observed in Figure 4, does not have income in the top 10 important features to predict borrower default or lateness.
- The grade of the loan seemed to have a really strong downwards trend with late or default borrowing percentage, as seen in 5. However, that did not show up in our model. That is probably because the grade is associated with other features in our dataset. Moreover, it could be a combination of loan amount and payment plan, which we suspect it is. The model chose to pick the raw features, which benefits the decision maker. That means that this loan classification could be redundant and can be removed in future loans.
- Jobs in the fields of driving or gambling still seem to have a higher probability of defaulting or paying late. Unfortunately, our moral model could not deny this claim. That is because neither the holistic model or the moral model could accommodate for the job title in their numerical algo- rithms.

For future work, we would like to achieve ~~a~~ higher overall accuracy. With more computing power, we could probably explore more of the hyperparameter space of boosting. We could also incorporate something about job title to see if our model would also refute item number three in the above list. Possibly a one-hot feature of the top ten job titles that have the highest probability of being late could be incorporated into our training dataset.

6 References

Wendy Kan. (2019 March). Lending Club Loan Data, Version 1. Retrieved April 2019 from https://www.kaggle.com/wendykan/lending-club-loan-data.