**DS-SF-33 Final Project Design Write Up and Approval**

**Project Problem and Hypothesis**

* What's the project about? What problem are you solving?

*The project is about using Lending Club’s latest loan payment information, credit score, credit bureau information and other demographic information to predict whether a borrower will make on time loan payment.*

* Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?

*The outcome or target variable is a binary discrete variable with 1 represent any type of late payment/collection and 0 represent on-time payment or paid off. This is a binary classification supervised learning problem.*

* What kind of impact do you think it could have?

*Solving this problem will help Lending Club identify potential risky borrowers and make corresponding adjustment on payment schedule, loan pricing and capital reserve strategy.*

* What do you think will have the most impact in predicting the value you are interested in solving for?

*Lending Club assigned loan grade, debt to income ratio, delinquency and utilization variables will be the most predictive features in predicting the default probability.*

**Datasets**

* Description of data set available, at the field level (see table)

|  |  |
| --- | --- |
| **Column Name** | **Column Description** |
| id | A unique LC assigned ID for the loan listing. |
| member\_id | A unique LC assigned Id for the borrower member. |
| loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| int\_rate | Interest Rate on the loan |
| installment | The monthly payment owed by the borrower if the loan originates. |
| grade | LC assigned loan grade |
| sub\_grade | LC assigned loan subgrade |
| emp\_title | The job title supplied by the Borrower when applying for the loan.\* |
| emp\_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| verification\_status | Indicates if the borrowers' income was verified by LC, not verified, or if the income source was verified |
| issue\_d | The month which the loan was funded |
| loan\_status | Current status of the loan |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| url | URL for the LC page with listing data. |
| desc | Loan description provided by the borrower |
| purpose | A category provided by the borrower for the loan request. |
| title | The loan title provided by the borrower |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| addr\_state | The state provided by the borrower in the loan application |
| dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_record | The number of months since the last public record. |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| pub\_rec | Number of derogatory public records |
| revol\_bal | Total credit revolving balance |
| revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| total\_pymnt | Payments received to date for total amount funded |
| total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| total\_rec\_prncp | Principal received to date |
| total\_rec\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |
| recoveries | post charge off gross recovery |
| collection\_recovery\_fee | post charge off collection fee |
| last\_pymnt\_d | Last month payment was received |
| last\_pymnt\_amnt | Last total payment amount received |
| next\_pymnt\_d | Next scheduled payment date |
| last\_credit\_pull\_d | The most recent month LC pulled credit for this loan |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income |
| verification\_status\_joint | Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified |
| acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| tot\_coll\_amt | Total collection amounts ever owed |
| tot\_cur\_bal | Total current balance of all accounts |
| open\_acc\_6m | Number of open trades in last 6 months |
| open\_il\_6m | Number of currently active installment trades |
| open\_il\_12m | Number of installment accounts opened in past 12 months |
| open\_il\_24m | Number of installment accounts opened in past 24 months |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| total\_bal\_il | Total current balance of all installment accounts |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| open\_rv\_12m | Number of revolving trades opened in past 12 months |
| open\_rv\_24m | Number of revolving trades opened in past 24 months |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| all\_util | Balance to credit limit on all trades |
| total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| inq\_fi | Number of personal finance inquiries |
| total\_cu\_tl | Number of finance trades |
| inq\_last\_12m | Number of credit inquiries in past 12 months |

* If from an API, include a sample return (this is usually included in API documentation!) (if doing this in markdown, use the javacription code tag)

**Domain knowledge**

* What experience do you already have around this area?

*I have more than 3 years of experience in credit risk modeling, spread across consumer credit risk modeling, Basel risk models and small business commercial credit risk modeling. I have experiences in model development, model validation, model monitoring and model deployment.*

* Does it relate or help inform the project in any way?

*It is closely related to the project and builds a solid business knowledge background for the problem.*

* What other research efforts exist?

*On Kaggle, multiple machine learning models have already been built by others to explore how to predict loan default using the information contained in the dataset, including but not limited to neutral networks, random forest and deep learning method using Tensor Flow.*

**Project Concerns**

* What questions do you have about your project? What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help).

*Don’t have many questions at this moment.*

* What are the assumptions and caveats to the problem?
  + What data do you not have access to but wish you had?

*FICO credit score used for loan underwriting, customer business/personal checking account, credit card statements data.*

* + What is already implied about the observations in your data set?

*Lending Club is a P2P lending company. Its customer base may be quite different from traditional financial institutions. I would imagine its overall credit default risk is higher since most of their customers could not seek credit from their primary financial institutions due to limited credit history, poor credit and much stricter underwriting policies. The learnings from this project can still be applied to the P2P lending industry.*

* What are the risks to the project?
  + What's the cost of your model being wrong? (What's the benefit of your model being right?)

*The model may create two types of errors. Type I error is when the borrower does not miss payment or go into default but the model predicts he/she will miss payment. This will increase unnecessary capital reserve for Lending Club which could have been used for more P2P loans and other business operations. Type II error is when the borrower is going to miss payment but the model predicts he/she will not. In that case, Lending Club will fail to set up enough capital reserve and face higher operating and compliance risk. For most financial institutions, type II error is relatively more important than type I.*

* + Is any of the data incorrect? Could it be incorrect?

*The data does not seem to be incorrect.*

**Outcomes**

* What do you expect the output to look like?

*The outcome will be a set of machine learning models output predicted default probabilities. I plan to use logistic regression with elastic net regularization, decision trees, random forest, stochastic gradient boosting tree and some other deep learning method available in Tensor Flow. Then I will generate model performance comparison chart/plot to compare key metrics like AUC, precision, F1 score of all these models.*

* What does your target audience expect the output to look like?

*My target audience will be risk analytics professionals, business partners and corporate legal team. They would expect to know exactly what features are being used by my model and relative importance of these features. In addition, model validations that justify the stability of model and potential business impact on company bottom lines.*

* What gain do you expect from your most important feature on its own?

Based on my experience, there will be a list of about 10 to 15 features in the final model. The top 3 features will have much larger marginal contributions to the overall model predictive power with the rest of the features helping capture other variations in the default probability that have not been explained by those 3 top features.

* How complicated does your model have to be?

*My final model can be a set of two models, one is a logistic regression with elastic net regularization, the other can be an ensemble deep learning method that potentially uses a combination of different kinds of machine learning models. The more complicated model will serve as benchmark to show case the best model performance that can be achieved using the data provided, while the simpler one is used for actual production and strategy creation, given the performance gap between them is not too wide.*

* How successful does your project have to be in order to be considered a "success"?

*Since credit score is not provided in this dataset, there is no benchmark score can be used to compare performance. But I should expect my outcome to make a difference in predicting default probabilities which means corresponding performance metrics have to be in certain bands like AUC should generally be greater than 0.75 and the model can do a great job in reducing type II errors (less misclassification errors in identifying bad loans as good loans).*

* What will you do if the project is a bust (this happens! but it shouldn't here)?

*It is possible that the project is a bust and no powerful predictive models can be created exhausting all possible machine learning algorithms. This may imply predictive features are not included in the dataset such as credit score. But at least, the model should be able to identify some top features and relative thresholds that can be used to create business rules in underwriting or loan portfolio management, which will definitely help the business to minimize loss on loan defaults.*