**Project Design Writeup**

* **Requirements:**
  + Well-articulated problem statement with "specific aim" and hypothesis, based on your lightning talk
    - **Problem Statement:**
      * + Using the loan data for all loans issued through the 2007-2015 in lending club, determine how likely the loan is approved and if approved, how much line to be assigned, based on the profile data and performance data
    - **Hypothesis:**
      * + The higher of the fico score, income, debt-to-income ration, the higher likelihood of being approved and higher line would be assigned.
  + An outline of any potential methods and models
    - **For predictor1: likelihood of being approved or not**
      * + Logistic Model
        + K Nearest Neighbors(KNN)
    - **For predictor2: line to be assigned**
      * + Linear Regression
  + Detailed explanation of extant data available (ie: build a data dictionary or link to pre-built data dictionaries)
    - **See attachment**
  + Describe any outstanding questions, assumptions, risks, caveats
    - **There are a mix of continuous and categorical independent variables**
    - **Some categorical variables, such as address, have many sub-level. Thus, hard to define if it is statistically significant**
    - **There are around 70 variables, how to reduce them**
    - **Check for the collinearity for those variables**
  + Demonstrate domain knowledge, including specific features or relevant benchmarks from similar projects
    - **Normally, risk score and previous payment behavior would be of most important**
    - **Cut-off rate for logistic model is the first 60 percent**
  + Define your goals and criteria, in order to explain what success looks like
    - **For predictor1: likelihood of being approved or not**
      * + Use the accuracy and AUC (area under the curve), The higher the accuracy the better. The bigger the AUC, the better.
    - **For predictor2: line to be assigned**
      * + Use the variance and bias (Mean Squared Error). The lower of the variance and bias, the better.
* **Bonus:**
  + Consider alternative hypotheses: if your project is a regression problem, is it possible to rewrite it as a classification problem?
    - **Yes. Create groups and transform them into category.**
  + "Convert" your goal metric from a statistical one (like Mean Squared Error) and tie it to something non-data people can understand, like a cost/benefit analysis, etc
    - **To generate the model that can generate the maximum profits while mitigate the default rate.**

Attachment:

|  |  |
| --- | --- |
| **LoanStatNew** | **Description** |
| addr\_state | The state provided by the borrower in the loan application |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| collection\_recovery\_fee | post charge off collection fee |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| desc | Loan description provided by the borrower |
| 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. |
| 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 |
| earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| 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. |
| emp\_title | The job title supplied by the Borrower when applying for the loan.\* |
| fico\_range\_high | The upper boundary range the borrower’s FICO at loan origination belongs to. |
| fico\_range\_low | The lower boundary range the borrower’s FICO at loan origination belongs to. |
| 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. |
| grade | LC assigned loan grade |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| id | A unique LC assigned ID for the loan listing. |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| is\_inc\_v | Indicates if income was verified by LC, not verified, or if the income source was verified |
| issue\_d | The month which the loan was funded |
| last\_credit\_pull\_d | The most recent month LC pulled credit for this loan |
| last\_fico\_range\_high | The upper boundary range the borrower’s last FICO pulled belongs to. |
| last\_fico\_range\_low | The lower boundary range the borrower’s last FICO pulled belongs to. |
| last\_pymnt\_amnt | Last total payment amount received |
| last\_pymnt\_d | Last month payment was received |
| 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. |
| loan\_status | Current status of the loan |
| member\_id | A unique LC assigned Id for the borrower member. |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| mths\_since\_last\_record | The number of months since the last public record. |
| next\_pymnt\_d | Next scheduled payment date |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| pub\_rec | Number of derogatory public records |
| purpose | A category provided by the borrower for the loan request. |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| recoveries | post charge off gross recovery |
| 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. |
| sub\_grade | LC assigned loan subgrade |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| title | The loan title provided by the borrower |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| 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\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |
| total\_rec\_prncp | Principal received to date |
| url | URL for the LC page with listing data. |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| 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 |
| 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 |
|  |  |
|  | \* Employer Title replaces Employer Name for all loans listed after 9/23/2013 |