upGrad

#LifeKoKaroLift

Lending Club Case Study

upGrad

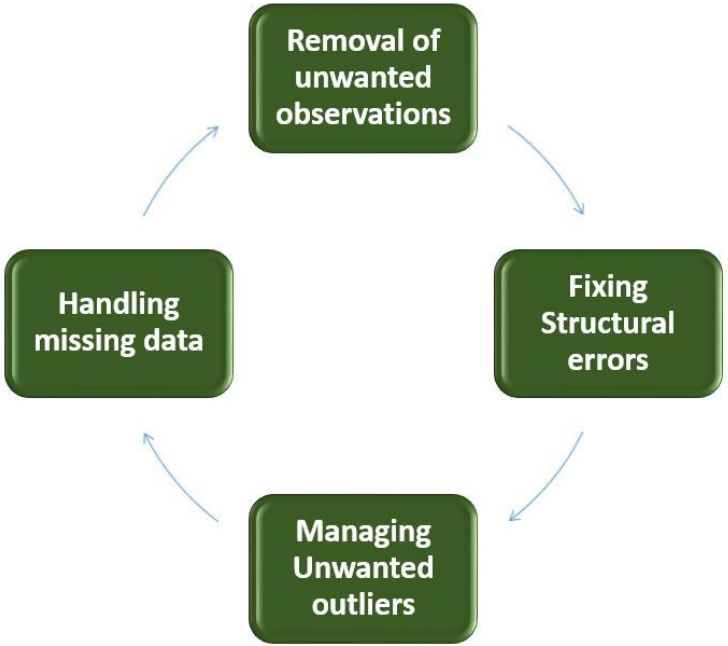
Contents

- Problem Statement
- Solution Approach
- Observations and Recommendations
- Appendix

Problem Statement

Company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Solution Approach : Overview



1. Using the data given which is about past loan applicants and whether they ‘defaulted’ or not. The aim is to identify patterns which indicate if a person is likely to default
2. As a next step we understood data given to us, where each record is information about a customer and each column defines his attributes
3. Next Step we performed was to clean data, the image on the left side summarizes the entire process as to how we arrived at a clean dataset, which was further used for analysis
4. After this we plotted necessary visuals and aggregated data at required levels to make some observations

Solution Approach : Data Cleaning

upGrad

Data Cleaning

1. Eliminated columns with a good percentage of missing values, as imputing the entire data would be equivalent to creating noise in data
2. Columns which contained random integers like member_id though it helps in identifying a customer uniquely, does not help us in our analysis and hence we chose to remove such columns
3. There are few columns which are absent during the loan application process, it can be post loan parameters or inputs from credit bureau's, we chose to remove these as well because they can never be qualified as predictors
4. There were also some repetitive columns like desc with same information being conveyed differently and hence considered using the purpose column which well classified the purpose of the loan
5. There were also text columns that required some complicated text mining to fix them and was also deemed not to be captured in the future, there is not point looking at this column as this data won't be available later
6. Removed single valued columns
7. Also eliminated rows where the customer was in the process of paying the loan as it does not fit the context, given its a imbalance dataset in terms of loan status, have been very careful while deleting rows
8. Examine every column from the remaining list, fix them make necessary transformations to make it useful, extract relevant data such as year, mnth etc also detected and removed outliers in income column

Solution Approach : List of Important Variables

upGrad

Important Variables after data cleaning :

- Loan Amount
- Issued Date
- Grade
- Sub Grade
- Annual Income
- House Ownership
- Term of the Loan
- Interest Rate for the loan
- Purpose of the loan
- Employment Tenure
- State, Pincode

Observations & Recommendations

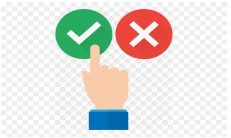
upGrad

Observations :

- Loans with higher amount are less in number and vice versa
- The term for all of the loans is either 3 or 5 years, popular one being 3 years
- There are lesser loans with higher interest rate and vice versa
- Grade B loans are common, double clicking on this A4 and B3 subgrade loans are the most popular
- Most of our customers are employed for more than 10 years
- Most of our customers either live in a Rented or a Mortgaged home (92%)
- Annual income is skewed distribution for our population of customers with median income at 59k$
- 56% of our applicants are either verified by LC or 3rd Party
- The most popular reason why our customers need loans is to consolidate their debts
- The company has been disbursing loans year on year in a incremental fashion and also observe this number to grow from Jan - Dec within a year
- Most of our loan applicants reside in California, New York, Florida, Texas and New Jersey

Observations & Recommendations

upGrad



Recommendations :

- Loans with higher amount are less in number and are more prone to be charged off, we should set a threshold for our loan values and install some more complex processes to approve such huge loans
- Longer termed loans are due to higher loan amount and are more prone to getting defaulted
- Charged off loans on average have higher interest rates compared to paid ones and hence we should continue to use the algorithm that is rightly assigning a higher interest rate to risky loans
- Also as observed, if the purpose of the loan is to set up small business, the recommendation here is to investigate into the business plan a bit and estimate profitability of the borrower's business idea, because we see them contributing to 8.5% of charged off loans
- Customers with Annual income below $40,000 are more likely to default, the recommendation is to check such customers even more thoroughly
- Verified customers are more likely to default which is counter intuitive, one reason being their DTI is high compared to those who don't default on the loans, also maybe the 3rd party whose verifying these customers need to reinvestigate their process as to why they are approving such faulty customers

(For any visualizations and data aggregations, please refer to the Jupyter Notebook)

upGrad

#LifeKoKaroLift