Assumption:

The following features are considered as critical features, 'id', 'loan\_amnt', 'term','int\_rate', 'sub\_grade', 'emp\_length','grade', 'annual\_inc', 'loan\_status', 'dti', 'mths\_since\_recent\_inq', 'revol\_util', 'bc\_open\_to\_buy', 'bc\_util', 'num\_op\_rev\_tl'

**loan\_amnt** — The listed amount of the loan applied by the borrower.

**term** — The number of payments on the loan, where values are in months and can be either 36 or 60.

**int\_rate** — The interest rate on the loan

**sub\_grade** — Assigned loan subgrade score based on borrower’s credit history

**emp\_length** — Borrow’s employment length in years.

**home\_ownership** — The homeownership status provided by the borrower (e.g., rent, own, mortgage, etc.)

**annual\_inc** — The self-reported annual income provided by the borrower

**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, divided by the borrower’s monthly income.

**mths\_since\_recent\_inq** — Months since most recent inquiry

**revol\_util** — Revolving line utilization rate, or the amount of credit the borrower uses relative to all available revolving credit.

**bc\_open\_to\_buy** — Total open to buy on revolving bankcards

**bc\_util** — Ratio of total current balance to high credit/credit limit for all bankcard accounts

**num\_op\_rev\_tl** — Number of open revolving accounts

**loan\_status** — Current loan status (e.g., fully paid or charged off). This is the label we are going to predict with the model.

Data Analysis Report for Loan Default Prediction

# Executive Summary

The data analysis report presents an analysis of loan training data to prepare for model training and loan default prediction. The goal is to help lenders better assess the risk of potential borrowers. The dataset `train\_data.csv` contains 145 features and 1827125 number of records.

The analysis used various statistical and machine learning techniques to identify the most significant predictors of loan default and to build a predictive model. Our analysis revealed that [insert key findings]. Based on these findings, we developed a predictive model that can accurately identify loans at risk of default. The model achieved an accuracy score of [insert score] on a test set, indicating its ability to generalize to new data.

# Data Cleaning and Preprocessing

Before conducting any analysis, we performed data cleaning and preprocessing to ensure that the data was of high quality and suitable for analysis. Our preprocessing steps included [insert steps], resulting in a clean dataset with [insert number] records.

## Assumption:

The following features are considered as critical features, 'id', 'loan\_amnt', 'term','int\_rate', 'sub\_grade', 'emp\_length','grade', 'annual\_inc', 'loan\_status', 'dti', 'mths\_since\_recent\_inq', 'revol\_util', 'bc\_open\_to\_buy', 'bc\_util', 'num\_op\_rev\_tl'

## Feature selection

151->108 (remove those over 90% missing rate)

After preselecting features, the next step is to deploy univariate analysis. This is when you analyze a single feature, and the most common two techniques you can use are 1) remove features with low variance (more than 90%) and 2) remove features that have a high amount of missing values.

#### Chi-square, p\_value, and Cramer’s V analyses for categorical variables

**identify highly correlated features**

There are a lot of ways to deal with it. The easiest way to detect highly correlated features is to use Pearson correlation and delete one of the perfectly (~ 90%) correlated features

from pandas\_profiling import ProfileReport   
profile = ProfileReport (loans, title = 'Loans Defaults Prediction', html = {'style': {'full\_width': True }}) profile

**Third, find a correlation between features and target variable**

profile = ProfileReport (trn, title = 'Loans Defaults Prediction', html = {'style': {'full\_width': True }})

profile

trn.corr() greater than 90% drop

# Exploratory Data Analysis

We conducted exploratory data analysis (EDA) to gain insights into the characteristics of the dataset and identify potential predictors of loan default. Our EDA included [insert methods], revealing [insert key findings].

# Feature Selection and Engineering

Based on our EDA, we selected [insert number] features for our predictive model. We also performed feature engineering to create new features that could improve the predictive power of our model. Our feature engineering steps included [insert steps], resulting in [insert number] features.

# Model Selection and Evaluation

We trained and evaluated several machine learning models to identify the most effective algorithm for predicting loan default. Our models included [insert models], with [insert best model] achieving the highest accuracy score of [insert score]. We also evaluated our model's performance using [insert metrics], achieving [insert performance].

# Conclusion

Our analysis demonstrates that it is possible to predict loan default with a high degree of accuracy using machine learning techniques. Our model, trained on a clean dataset with carefully selected features, achieved an accuracy score of [insert score], indicating its potential for real-world application. Further research could explore additional features and models to improve the predictive power of our approach.

# Reference

<https://medium.com/@data.science.enthusiast/feature-selection-techniques-forward-backward-wrapper-selection-9587f3c70cfa>

Chi-square statistical test:

<https://towardsdatascience.com/levels-of-measurement-statistics-and-python-implementations-8ff8e7867d0b#:~:text=Chi%2Dsquare&text=It%20is%20a%20test%20to,the%20strength%20of%20the%20correlation.&text=As%20the%20sample%20size%20increases,proportion%20of%20the%20expected%20value>.