CREDIT RISK PREDICTION

By Fitria Dwi Wulandari

TABLE OF CONTENTS

01 PROBLEM RESEARCH

02 DATA PROCESSING

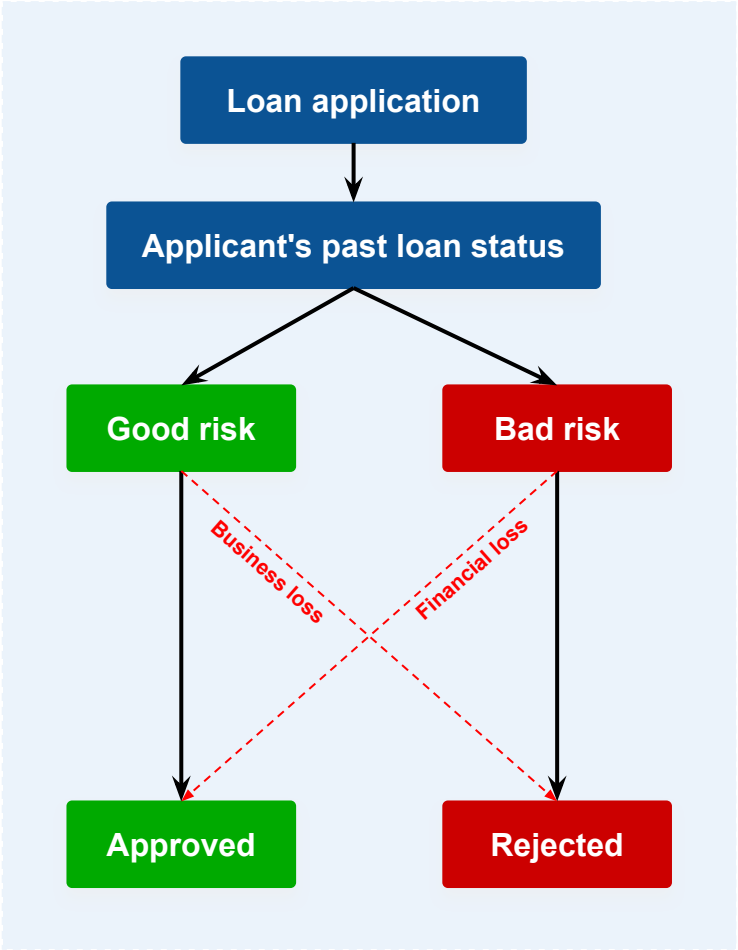
03 BUSINESS INSIGHTS

04 DATA MODELING

05 BUSINESS RECOMMENDATION

01

PROBLEM RESEARCH



BUSINESS UNDERSTANDING

The data contains the information about past loans of applicants and whether they labeled as a good risk or not. When a applicant applies for a loan, there are two type of risks, namely:

1. **Good risk** consists of **Fully Paid**, **Current**, and **In Grace Period**. Applicants with this label are more likely to get their loan approved in the future.
2. **Bad Risk** consists of **Late**, **Default**, and **Charged Off**. Applicants with this label are unlikely to get their loan approved in the future.

PROBLEM STATEMENT

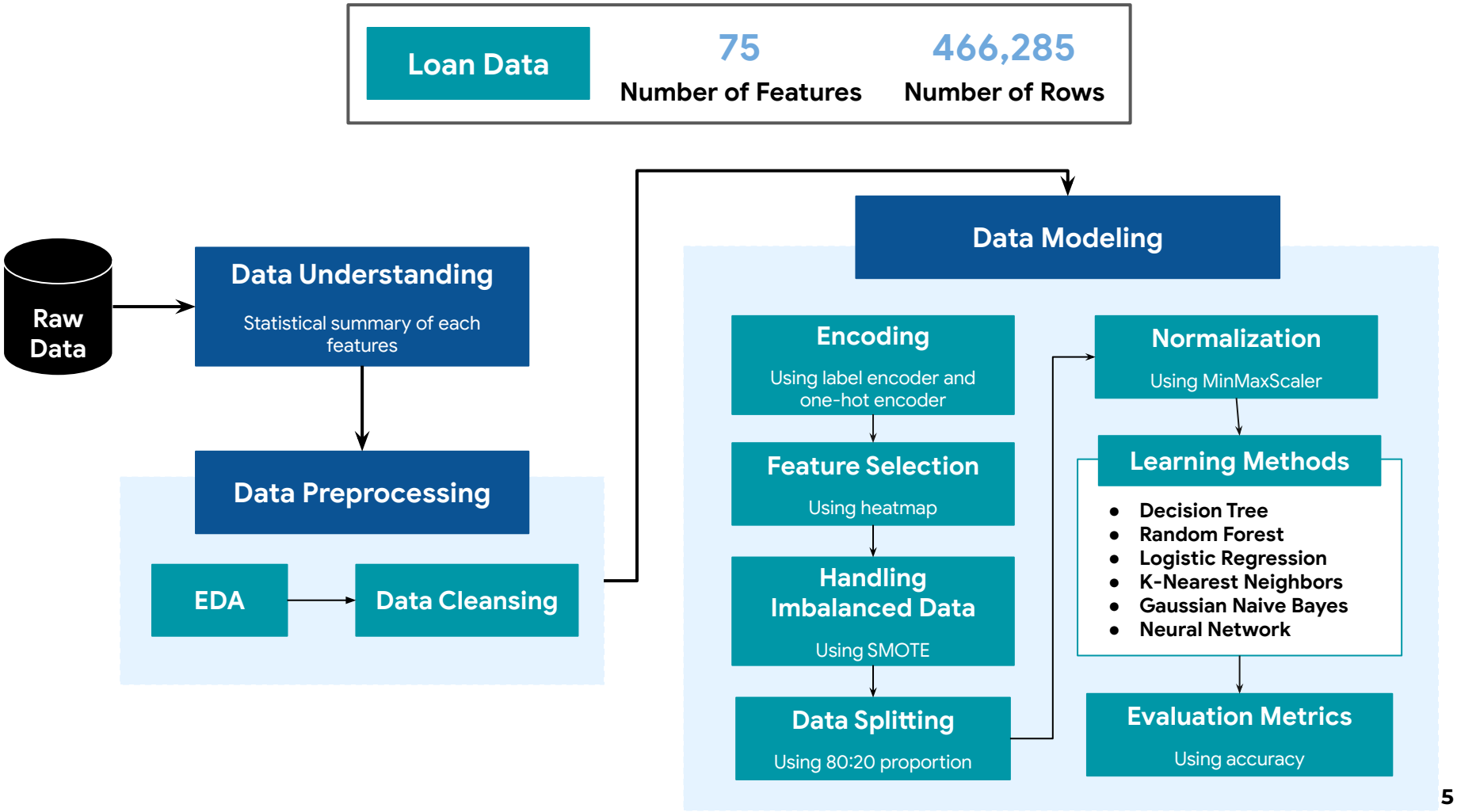
Lending loans to ‘bad risk’ applicants is the largest source of financial loss. Credit loss is the amount of money lost by the lender when the applicant refuses to pay or runs away with the money owed.

BUSINESS OBJECTIVES

1. Identify patterns that indicate if a person is unlikely to repay the loan or labeled as a bad risk.
2. Implement machine learning algorithms to build a predictive model to predict loan risk from applicants.

02

DATA PROCESSING


Loan Data
75
Number of Features
466,285
Number of Rows
Raw Data
Data Understanding
Statistical summary of each features
Data Preprocessing
EDA
Data Cleansing
Data Modeling
Encoding
Using label encoder and one-hot encoder
Feature Selection
Using heatmap
Handling Imbalanced Data
Using SMOTE
Data Splitting
Using 80:20 proportion
Normalization
Using MinMaxScaler
Learning Methods
Decision Tree
Random Forest
Logistic Regression
K-Nearest Neighbors
Gaussian Naive Bayes
Neural Network
Evaluation Metrics
Using accuracy

03

BUSINESS INSIGHTS

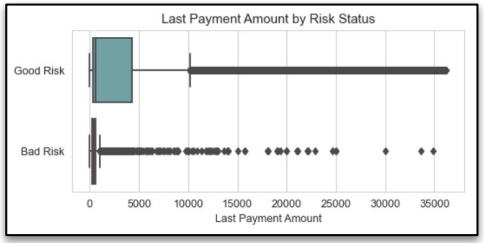
TOTAL LOSS SUFFERED BY THE COMPANY

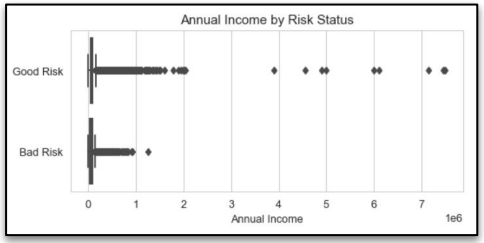
Loan Status	Total Loss	% of Total Loss	Total Applicants	Average Loss
Charged Off	$574,356,330	83.49%	43,236	$13,284
Late	$102,293,296	14.87%	8,118	$12,600
Default	$11,299,446	1.64%	832	$13,581

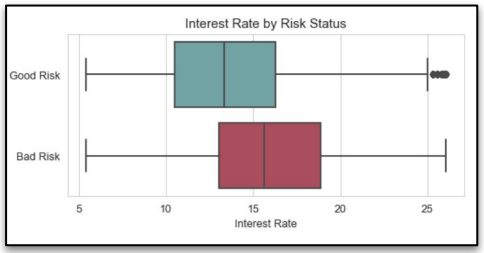
The loan status of **CHARGED OFF** is the **biggest source of loss** (**83%**) for the company with a total loss of 574 million from 43,236 applicants.

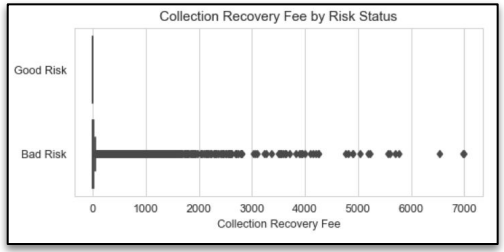
The loan status of **LATE** contributed **14%** to the company's losses with a total loss of 102 million from 8,118 applicants.

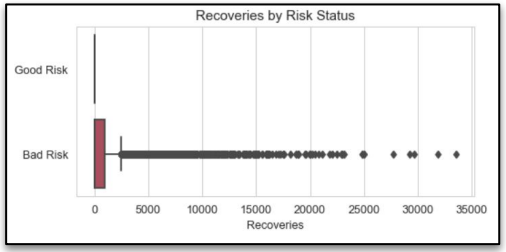
- The loan status of **DEFAULT** only contributed **1.64%** to the company's losses with a total loss of 11 million from 832 applicants.
- But this loan status has the **highest average loss value** of $13,581 per applicant.











Applicants with a **low last payment amount** are more likely not to repay their loans.

Applicants with **low annual incomes** are more likely not to repay their loans.

Applicants with the **high interest rate** have a high chance of **not being able to repay** the loan.

Applicants with **recoveries value** and **collection recovery fee greater than 0** are most likely not to repay their loan.

04

DATA MODELING

MODEL COMPARISON

Algorithms	Training Accuracy Score	Testing Accuracy Score	Error Margin
Decision Tree	99.98%	98.32%	1.66%
Random Forest	99.98%	99.11%	0.87%
Logistic Regression	88.38%	88.41%	0.03%
K-Nearest Neighbor	97.61%	96.7%	0.91%
Gaussian Naive Bayes	75.08%	74.98%	0.1%
Neural Network	98.99%	98.72%	0.27%
XGBoost Classifier	99.39%	99.08%	0.31%
Gradient Boosting Classifier	97.56%	97.52%	0.04%

- The best model to predict the risk status of loan applications is **Random Forest.**
- The difference in accuracy between training and testing sets on the random forest model is smaller than in the decision tree model.
- Although, the XGBoost classifier also has very high accuracy values, and the difference is smaller than the random forest model. The random forest model seems not only to perform better on the training set but also the testing set. That means the random forest model has a better generalization than the XGBoost classifier model.

BEST MODEL: RANDOM FOREST CLASSIFIER

Features

The model only consists of **17** features that are correlated with risk status compared to other features.

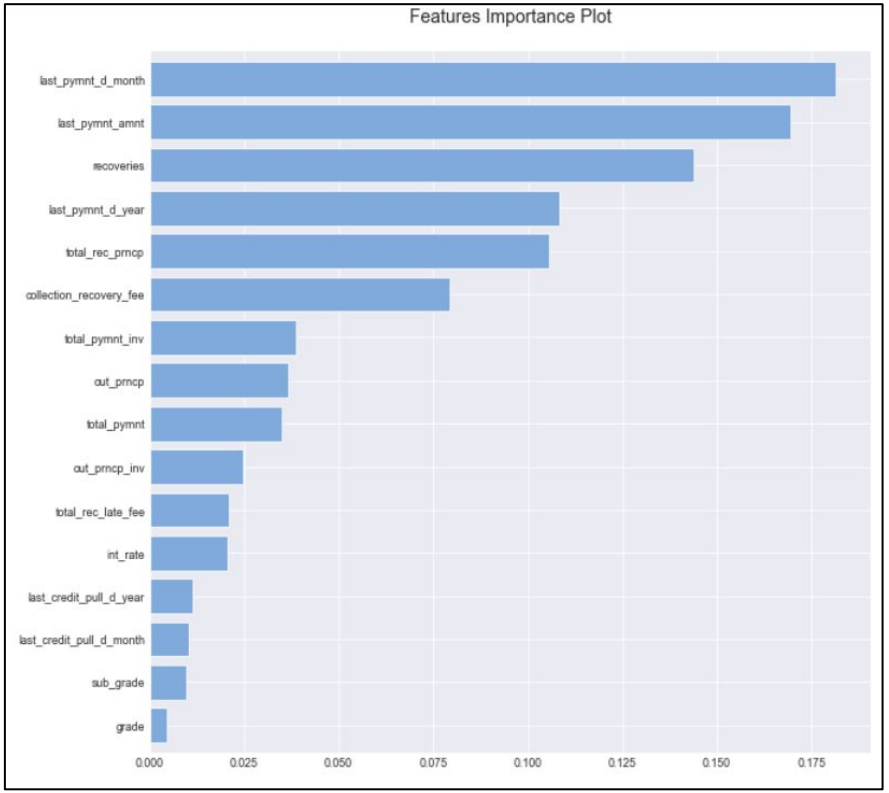
Performance

Random forest model gives **99.11%** correct results on testing data.

There is **0.87%** error margin

Top 5 important features

- Last payment month
- Last payment amount
- Total principal received
- Recoveries
- Last payment year



05

BUSINESS RECOMMENDATION

RECOMMENDATION

- **Last payment month**, **last payment amount, recoveries value**, the **total principal received**, and **last payment year** is the most important features to identify whether the applicant has the possibility of not repaying the loan. The company needs to monitor these indicators to reduce the risk of loss.
- In the future, If there are applicants with those indicators, then the company can take action such as rejecting their loan, reducing the amount of the loan, or lending at a higher interest rate to avoid and reduce the total loss suffered by the company.

You can see the entire project documentation here!

https://github.com/fitria-dwi/Credit-Risk-Prediction

THANK YOU