CREDIT RISK ANALYSIS

COMPANY PROFILE PRESENTATION

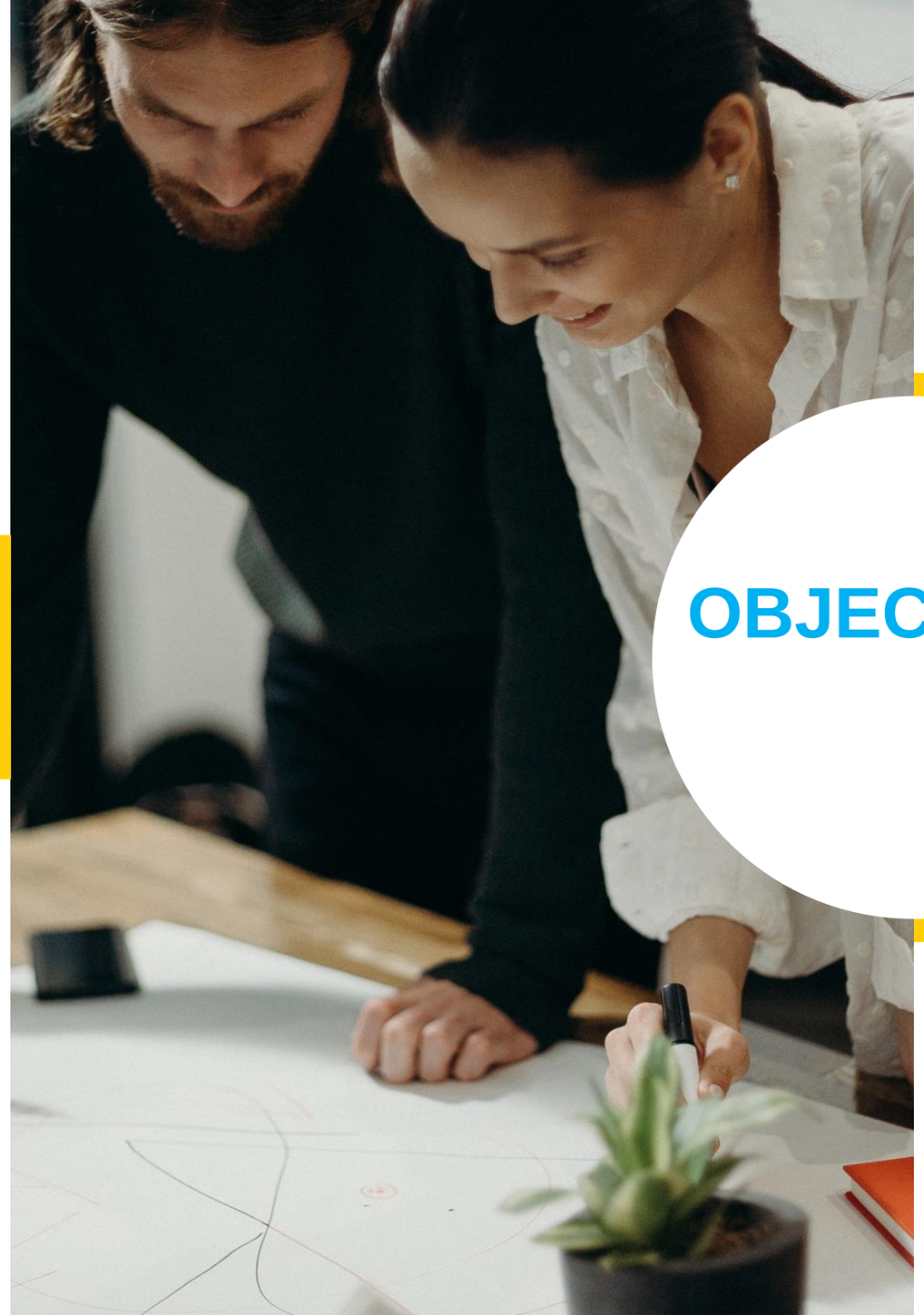
Table of Content

OVERVIEW

Having good credit scores and ratings raises ones eligibility for high quality loans on very agreeable terms. Good financial literacy dictates no living beyond whatever you cannot afford.

Loan and asset recovery departments often fail to meet their targets due to high default rates. This underscores the accuracy of determining the credit-worthiness of clients..

In this age of Financial Technology, machine learning plays a very crucial role in the credit analysis process. Many companies leverage the power of Machine learning algorithms to ensure a clear and precise credit screening process..



OBJECTIVES

After being approached by a Fintech start up who are said to have very high default rates, we set out to:

1. Find the income and defaults distribution across states and cities.

2. Find out the optimum loan credit score below which one is not available for a loan.

3. Develop a classification model for the Fintech to help in prediction of loan defaulters .

DATA

SOURCE

Data to analyze customer behavior and lifestyle to determine their credit worthiness scores. Data was sourced from https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior.

Description

The data had a mix of numerical and integer data types. The City and States columns had a very high unique values. (High Cardinality)

DATA (Columns)

6

Total number of Object Data types

11

Total number Of columns

252,000

Total Number of Columns

0

No Missing Values

Int. Cols

- Income Rates
- Job Experience
- Job & Hse yrs

Objects

- House ownership
- Car Ownership

DATA (Columns)

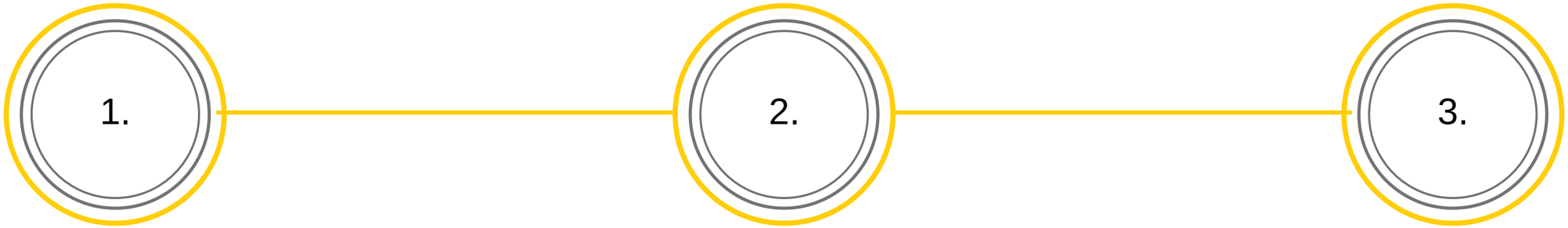
- Marital Status
- Hse & Car ownership

Objects

- City
- State

Feature Engineering

Curse of cardinality



1. The CITY column had more than 350 unique values with very high value counts across states.
2. The profession column had more than 50 unique values.
3. The state column also had many unique values though less than 50.

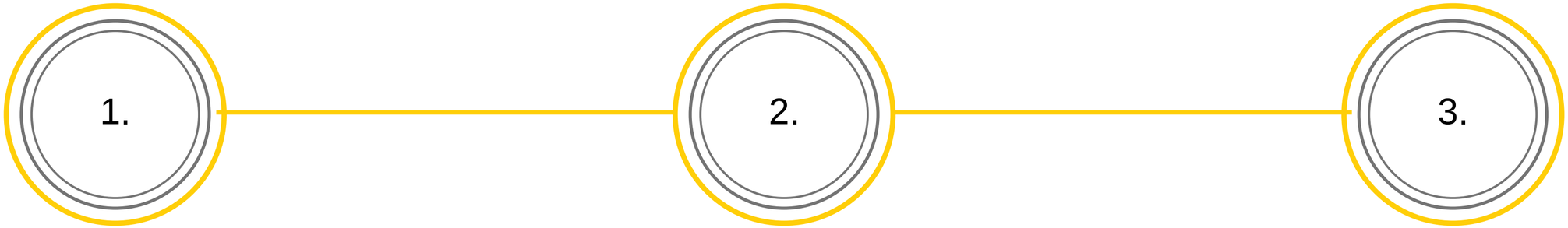
One Hot encoding these columns would result to memory issues.

Label encoding them would also result to loss of high negative correlations.

The technique of encoding by their value counts known as Count Encoding was very handy.

Values are replaced by their counts to maintain the correlation.

Data Analysis



Univariate Analysis.

1. Individual value counts and distributions.
(Count & Bar plots)

Bivariate Analysis

1. Age distributions across professions.
2. Income distributions across states, professions and families
3. Defaulters Vs. Non defaulters distributions.
(Box plots & bar plots)

Multivariate Analysis

1. Pair plots
2. Group by plots.

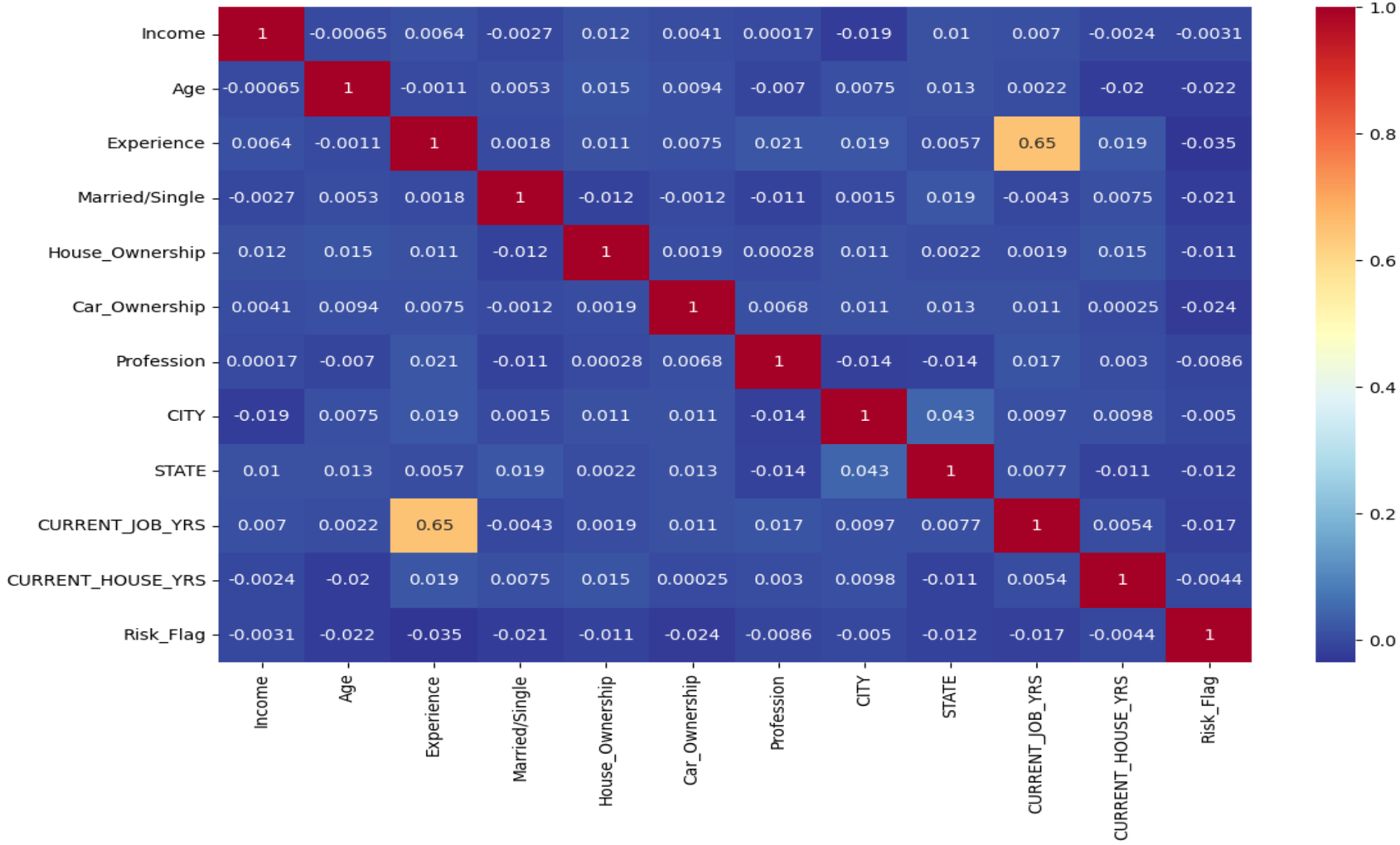
Heatmap

Highly Correlated features

The **Experience** and **CURRENT job years** columns were highly correlated in the sense that as job years increase, so does the Experience.

The Job years column was dropped during modelling to prevent multicollinearity.

All the other features had negative correlations with the target column.



Modelling

Bagging Classifier

Ensemble classifier that handles class imbalance well but must have a base estimator.
(Biased towards majority class)

Random Forest Classifier

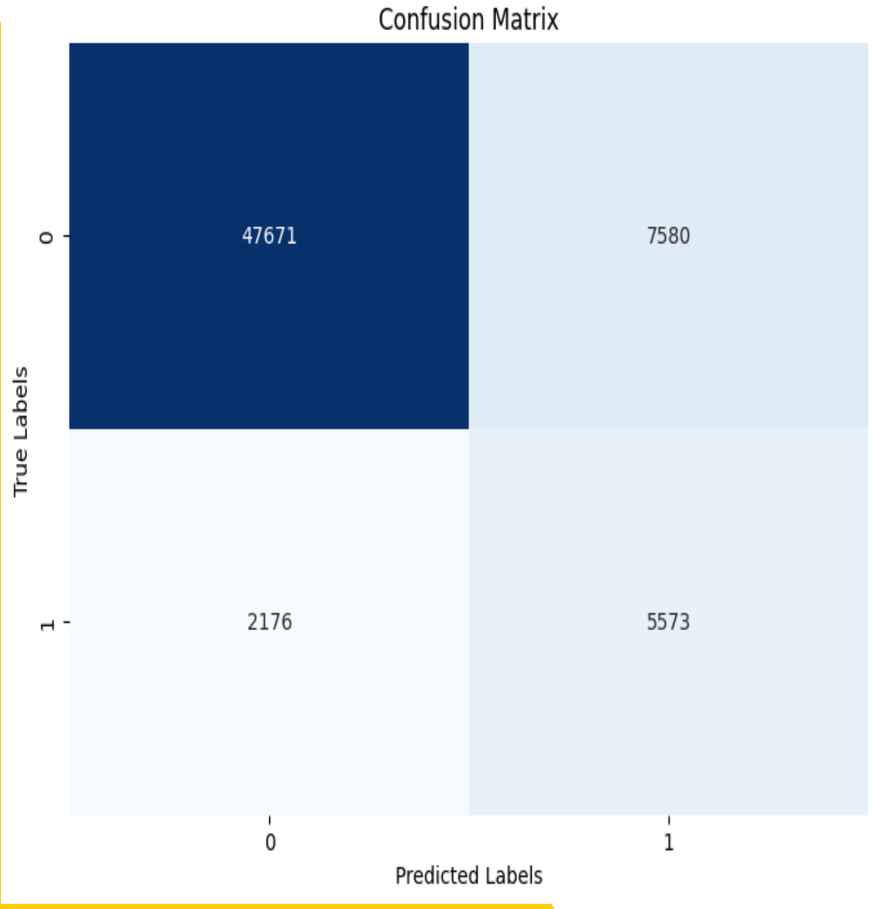
Ensemble Classifier
(Class imbalance)

XGBoost Classifier

Gradient Boosting algorithm with fast training speeds and good accuracies.
(Biased towards majority class)

CatBoost Classifier

Gradient boosting algorithm built for classification problems with very good handling of large datasets.
(gives more priority to the minority class)



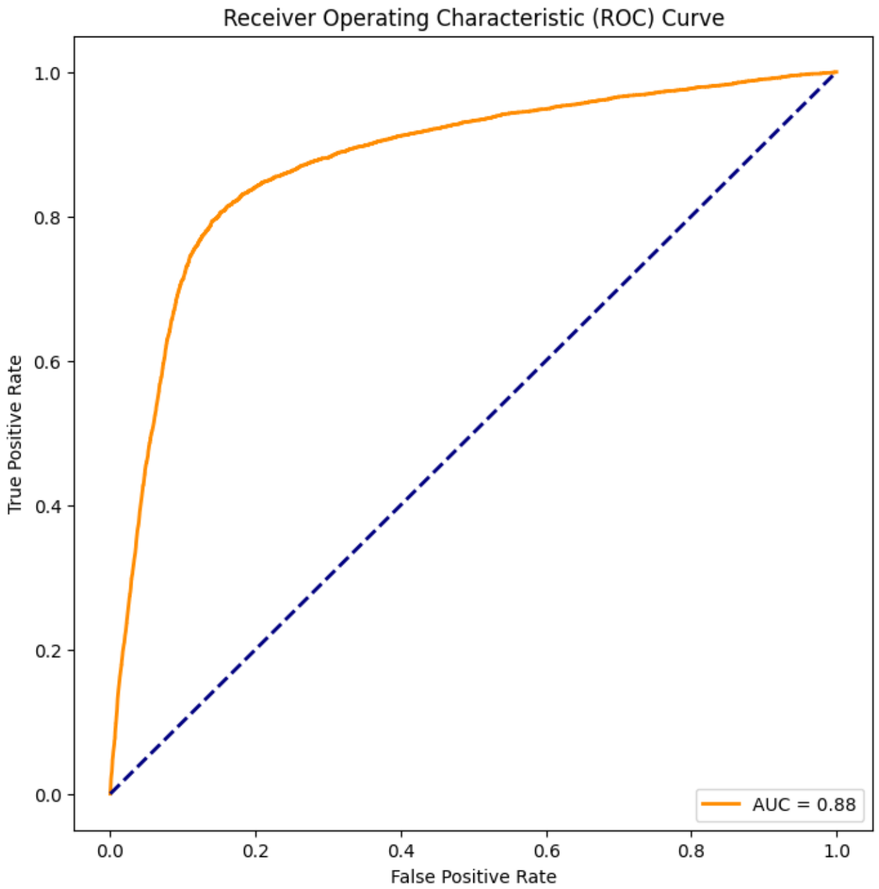
CATBOOST CLASSIFIER

I chose this classifier because:

1. Fast training speeds.
2. Handles categorical variables well.
3. Optimizes the log loss function used in classification problems.
4. Handles class imbalances well.

This classifier handles the **False negatives** pretty well and does not get biased to the majority class. The false negatives are bad for business and as such it is necessary to reduce them as possible.

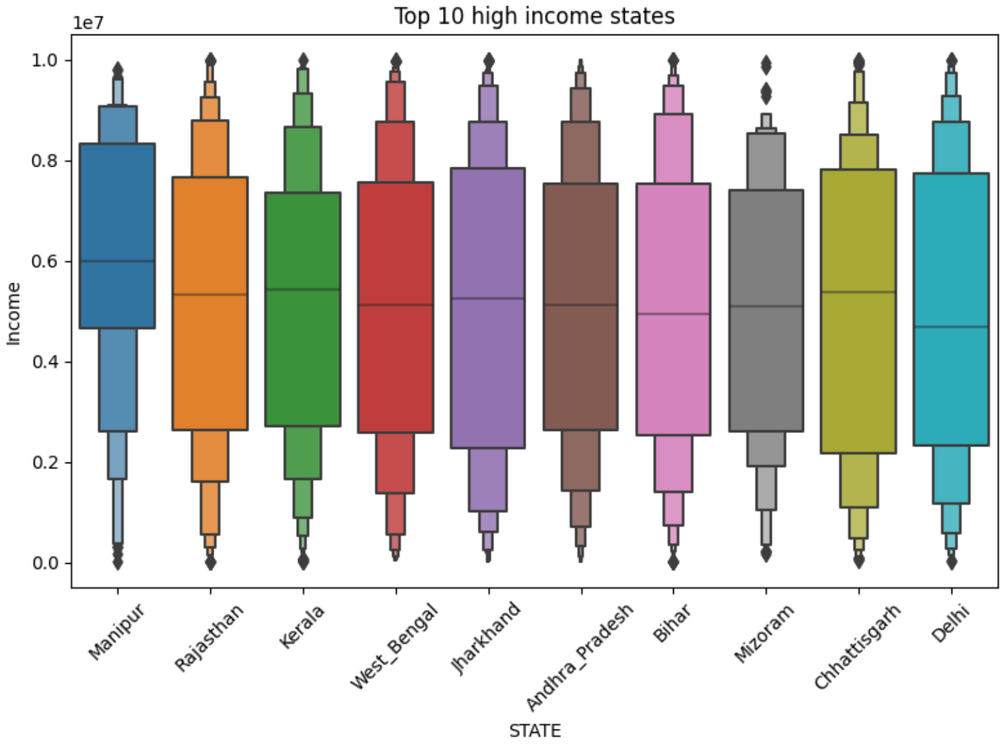
False Negatives are those people who have defaulted but then have been predicted as non-defaulters. This will lead to huge losses because if the number is high there won't be any follow-ups by the loan department.



Reccommendations.1

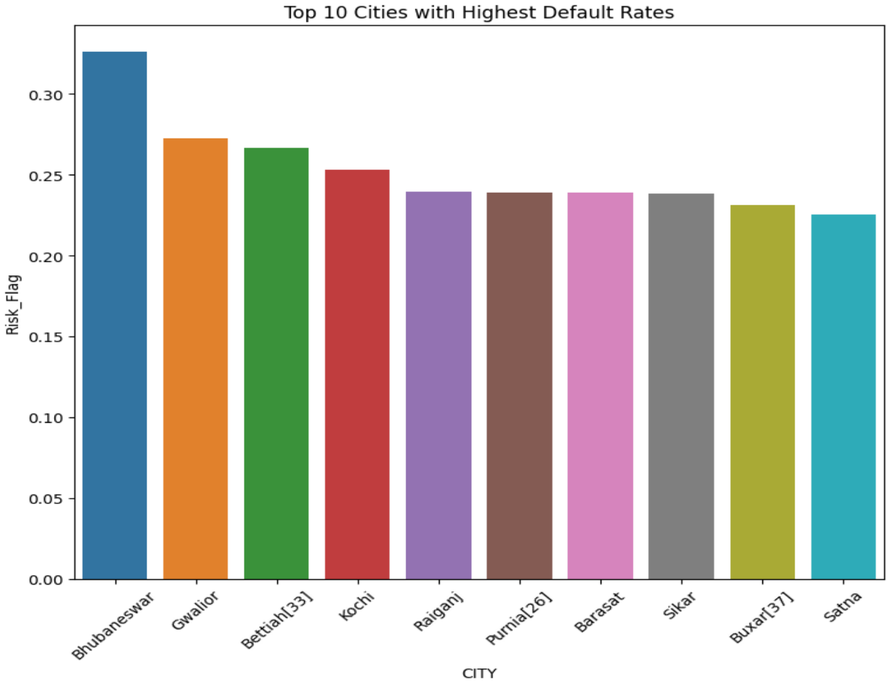
The Microfinance company should consider setting their loan credit score at 0.6 because this is the optimal point where the False Positive rate is reduced and the True positive rate is high.

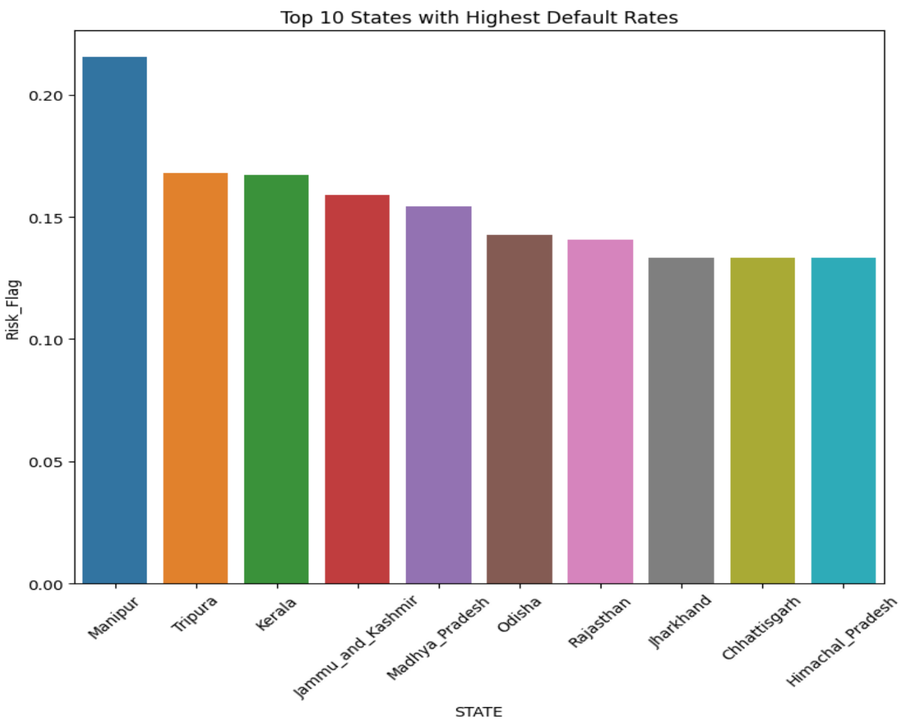
Clients who attain a credit score below o.6 should not be eligible for loans.



Reccommendations.2

The Microfinance group should consider venturing into high end income areas where the default rates may be significantly lower as compared to other areas.





Reccommendations.3

The head of loan department should consider re-assessing the credit policies in the cities of Gwalior, Bhubaneswar and Bettiah.

The cities Manipur, Tripura and Karaia also need stringent policies in the loan eligibility and clearance process. The loan eligibility cut point should never be compromised.

1. Data provided next time should not have high cardinality.

2. The loan department should asses how well loan repayments perform at different set values. Start at 0.6 and asses how well the loans are repaid at that rate.

THANK YOU

+123-456-7890

Microfinance @gmail.com

www.microfinanceworld.com