PREDICTING LOAN DEFAULT

SHASHI BHUSHAN
SPRING 2024
DSC 680

AGENDA

Introduction

Data Selection

Modeling and Methods Used

Interpretation of Analysis /
Model Results

Conclusion

INTRODUCTION

INTRODUCTION

- In 3Q 2021, 64 million Americans had debt with collection agencies
- In 2008 financial crisis, mortgage default rates increased significantly

DATA SELECTION

KAGGLE

- Resource Link: https://www.kaggle.com/datasets/yasserh/loan-default-dataset/data
- Loan default noted as 0 or 1
- Data collected in 2022
- Both categorical and numerical data
- 148,670 rows

Features - ID, year, loan limit, gender, approved in advance, loan type, loan purpose, credit worthiness, open credit, business or commercial, loan amount, interest rate, upfront charges, loan term, negative amortization, lump sum payment, property value, construction type, occupancy type, secured by, total units, income, credit type, credit score, co-application credit type, age group, submission of application to institution or not, LTV, region, security type, Status, debt to income ratio.

MODELS AND METHODS USED

VISUALIZATIONS

Boxplots

1. Loan amount with categories positive loan default (1) and negative loan default (0)
2. Income with categories positive loan default (1) and negative loan default (0)
3. Credit Score with categories positive loan default (1) and negative loan default (0)

Bar charts

For positive default cases with below categorical variables:

- Gender
- Loan purpose
- Credit worthiness
- Age
- Region
- Security type

DATA PREPARATION

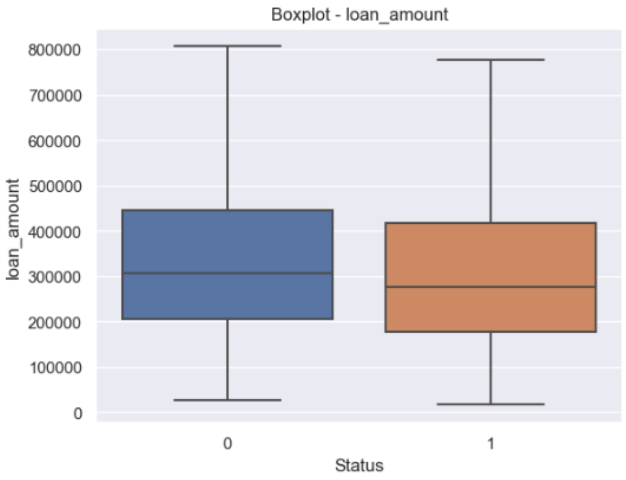
- Variables with > 10% missing values dropped
- Missing values imputed as noted below:
  - Categorical variables – mode
  - Numerical variables – KNN imputer using 5 neighbors
- Dummy variables created – one hot encoding

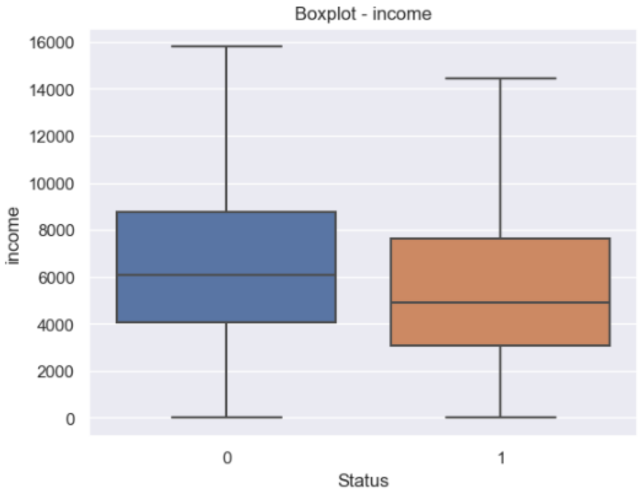
MODELING

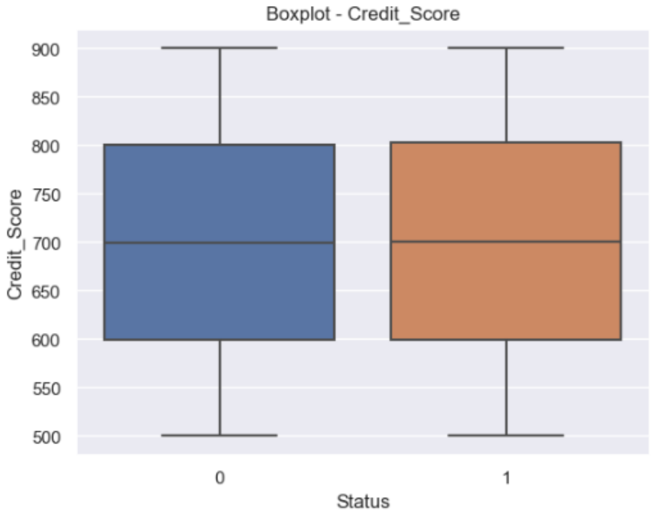
- Target outcome – 1 or 0
- Created two models:
  - Random Forest
  - Naïve Bayes

INTERPRETATION OF ANALYSIS / MODEL RESULTS

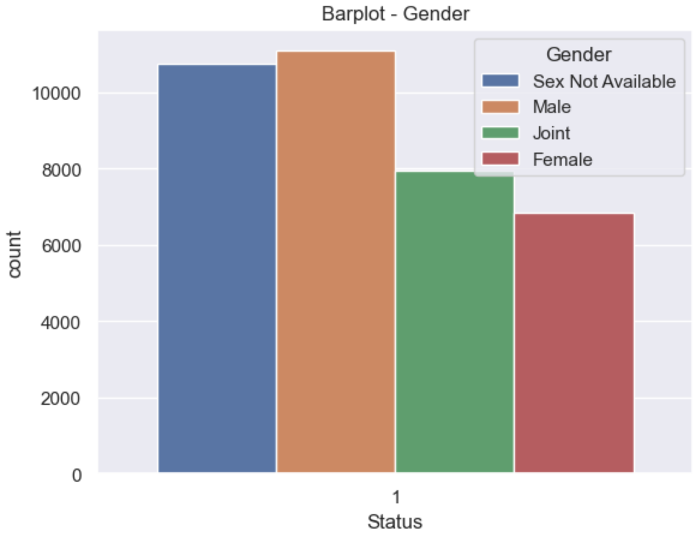
VISUALIZATIONS

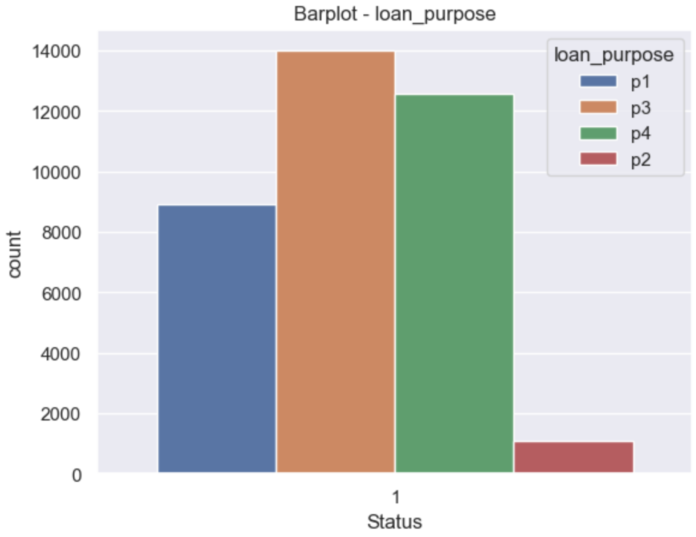


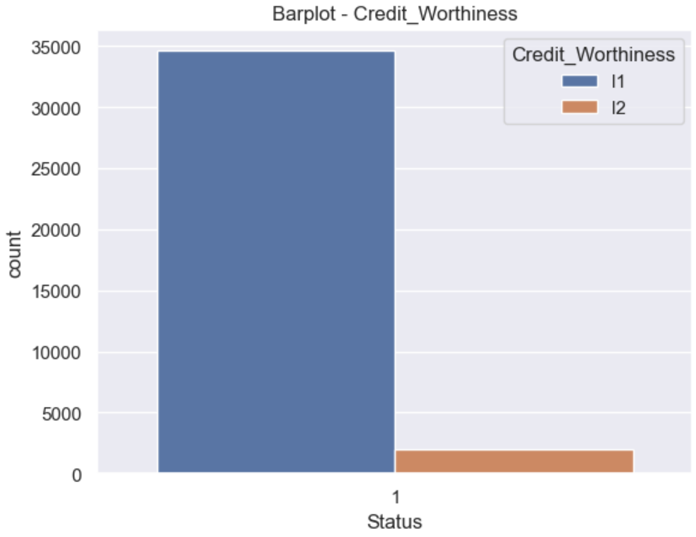




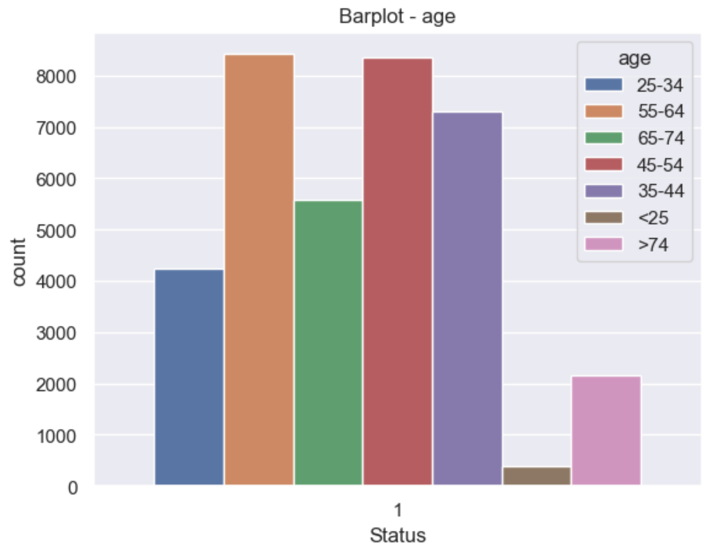
VISUALIZATIONS

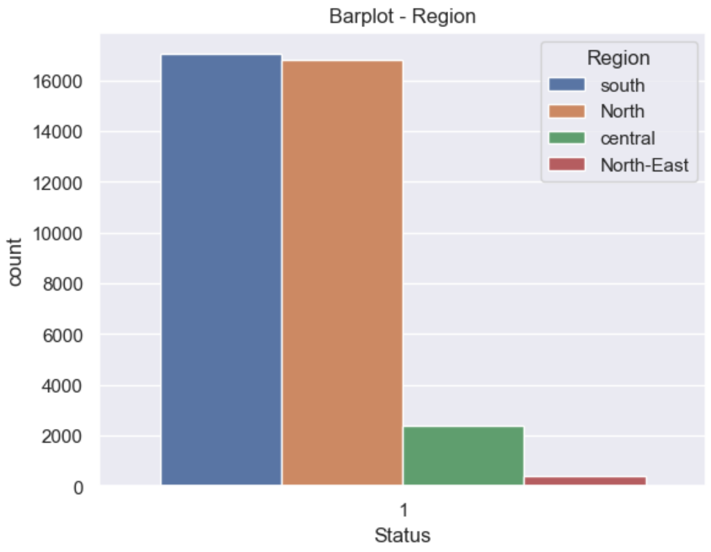


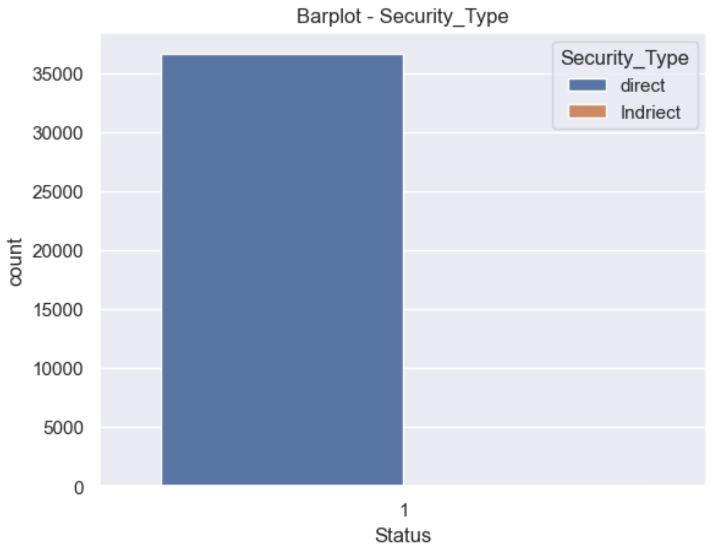




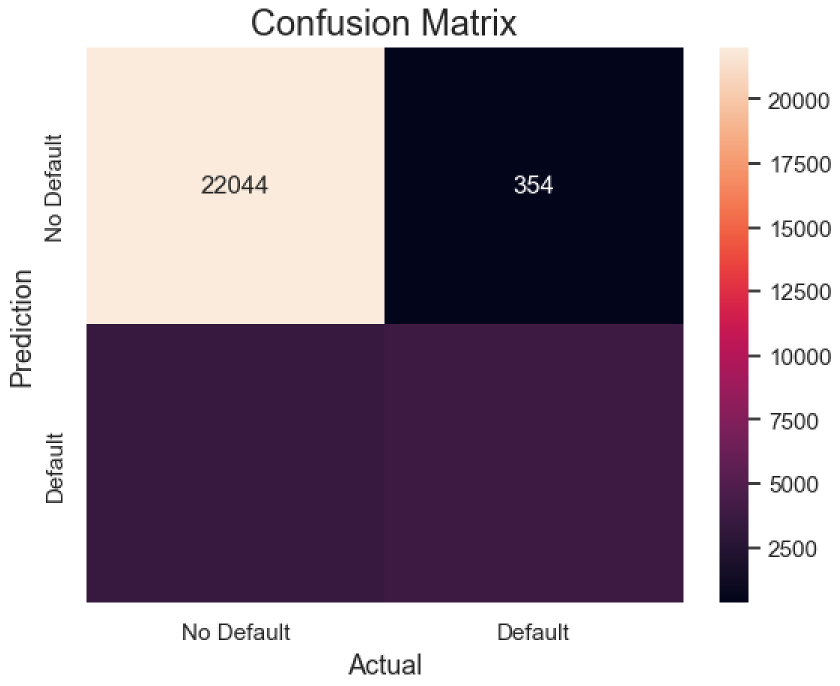
VISUALIZATIONS





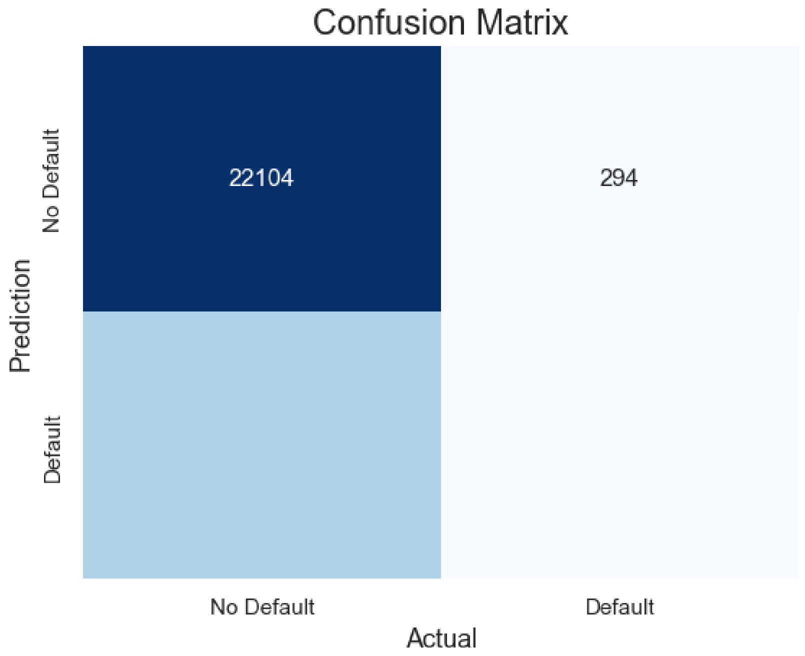


MODEL RESULT INTERPRETATION – RANDOM FOREST



	precision	recall	f1-score	support
0	0.86	0.98	0.92	22398
1	0.91	0.52	0.66	7336
accuracy			0.87	29734
macro avg	0.89	0.75	0.79	29734
weighted avg	0.88	0.87	0.86	29734

MODEL RESULT INTERPRETATION – NAÏVE BAYES

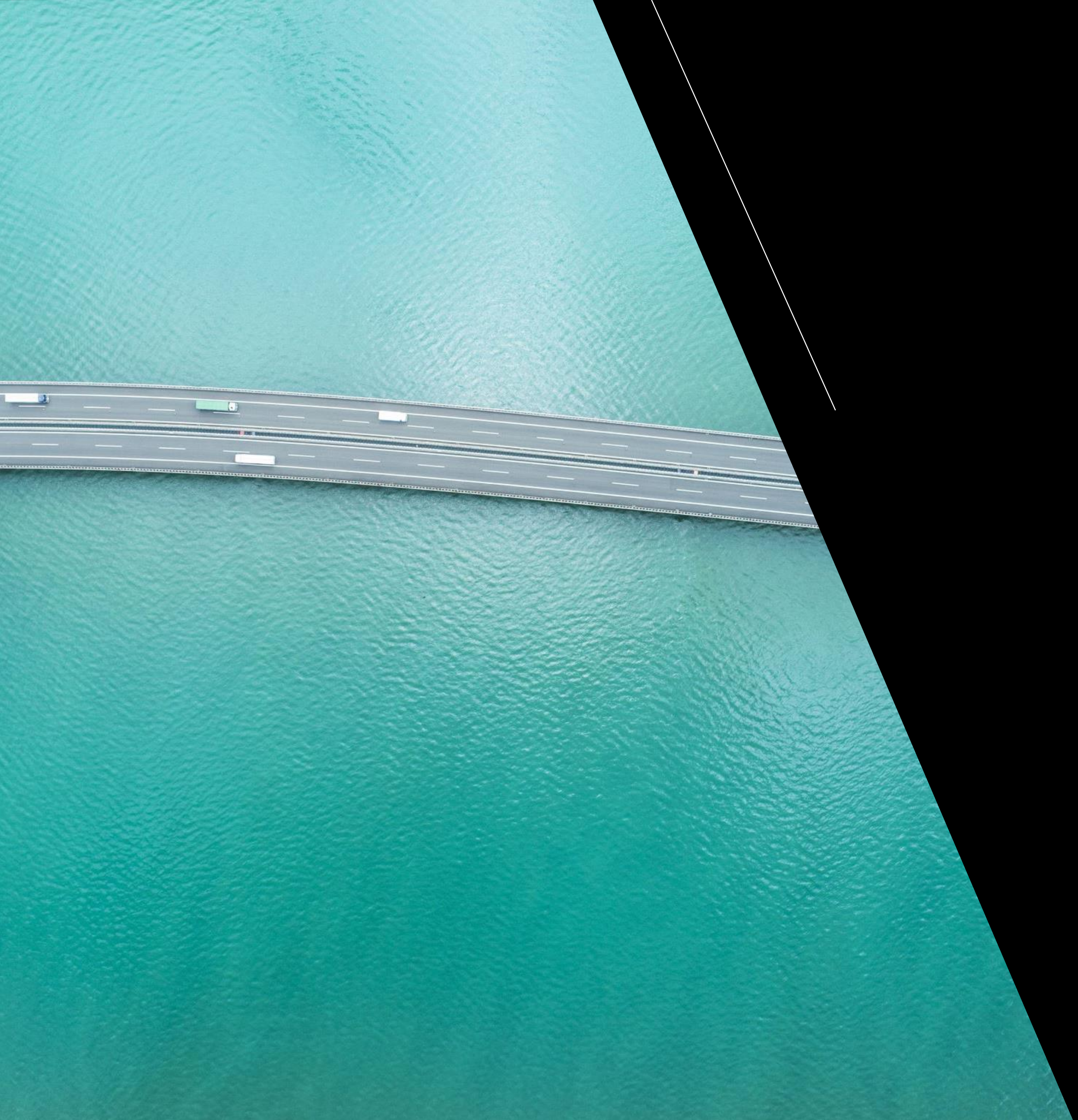


	precision	recall	f1-score	support
0	0.76	0.99	0.86	22398
1	0.37	0.02	0.05	7336
accuracy			0.75	29734
macro avg	0.56	0.51	0.45	29734
weighted avg	0.66	0.75	0.66	29734

CONCLUSION

Recommend Random Forest Model

- Higher accuracy for both 0 and 1
- High Precision, recall and f1 scores



THANK YOU