Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Project 2 Milestone 3

Business Problem

Loan default is defined as instances classified as such when borrowers consistently miss monthly payments for an extended period (typically 90 days or more). In the recent inflationary environment loan default cases have seemed to experience an uptick.

Banks or lending agencies have to take a lot of risk when approving loan applications given the long duration of repayment in case of home loans running into 15 to 30 years. Collection process is incredibly challenging for the families and attribute to emotional stress, worsens the financial strain, affects family relationships, and negatively impact children within households. In addition, loan defaults negatively impact the financial performance of the financial institutions.

This clearly represents a problem that requires to be solved and a solution is implemented where cases of defaults are minimized.

Background / History

In the third quarter of 2021, approximately sixty-four million Americans had one or more debt with collection agencies indicating default status (Carther, 2022). In 2008 financial crisis, mortgage default rates had increased significantly due to approval of loans to people who were unable to pay on sustained basis. This period highlighted the need for a robust approval process so that the lenders can minimize the losses.

Data Explanation

Source - https://www.kaggle.com/datasets/yasserh/loan-default-dataset/data.

Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

The dataset has thirty-four variables with status as the target variable noted as Y or N. It has 148,670 instances in total containing a mix of data type (categorical and numerical).

Features include 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.

These features should be sufficient to train a model for the binary classification problem of predicting potential default.

Data Preparation

I checked for the missing value in the dataset and found that rate of interest, Interest rate spread, upfront charges, property value, LTV, and dtir1 had more than 10% of values missing. Therefore, I dropped these features from the dataset. ID and year were also dropped as identifier would not add value to the model and year only had 2019 as values. I checked for duplicate rows and found none. Missing values in the categorical variables were imputed using mode and then dummy variables were created using one hot encoding and removing redundant complementary variables created. Numerical variables were imputed using KNN imputer using 5 neighbors.

Heatmap was created to check correlation between features and removed secured by land as it had perfect negative correlation with security type direct.

Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Methods

Visualizations

I used the following visualizations to explain my data:

- Boxplots
    - Loan amount with categories positive loan default (1) and negative loan default (0) (Figure 1)
    - Income with categories positive loan default (1) and negative loan default (0) (Figure 2)
    - Credit Score with categories positive loan default (1) and negative loan default (0) (Figure 3)
- Following bar charts for positive default cases to see how each category within below categorical features stacked up to show any trend.
    - Gender
    - Loan purpose
    - Credit worthiness
    - Age
    - Region
    - Security type

Analysis / Modeling:

Our target variable, outcome (0 or 1) is binary. I built and evaluated two models namely Random Forest classifier and Naïve Bayesian classifier.

Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Random Forest is an ensemble learning method that combines multiple decision trees to make robust predictions (Wood, n.d.). In binary classification it can easily predict whether an input belongs to one of two classes (e.g., spam vs. not spam, fraud vs. non-fraud).

Gaussian Naive Bayes is a classification technique commonly used for binary classification problems. It assumes that each class follows a normal distribution.

Interpretation of Analysis / Model Results

Visualizations

1. Median loan amount for default cases was lower than that for non-default cases.
2. Median income level for default cases were lower. This appears to be counter intuitive.
3. Median credit scores were same for both non-default and default cases.
4. Male individuals are more in default. Female loanees were about 35% less and indicating inclusion of females in joint loanees driving down default cases within this category.
5. Loan purpose p3 had highest default cases.
6. Age groups 45-64 had highest default cases.
7. South and North default cases were found to be significantly highest than central and northeast regions.
8. Security type direct were the only category with default cases.

Modeling / Interpretation

Two models produced decent accuracy percentages with Random Forest and Naïve Bayesian producing ~87%%, and ~75%, respectively. However, upon generation of the classification reports for the two models, precision, recall, and f1-score for Random Forest model were much higher for prediction of default cases values being 91%, 52%, and 66%, respectively. (Figure 10)

Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Conclusion

Precision is the proportion of every observation predicted to be positive that is actually positive. Recall is the ratio of true positives to sum of true positives and false negatives. F1 score is the harmonic mean of precision and recall and provides a balanced number between precision and recall.

Since all the above values are higher for default cases for random Forest model, I recommend use of this model for predicting loan default cases.

Assumptions

I assumed that if the number of missing values were over 10% for a feature, it would negatively impact the model and removed from the analysis.

Limitations and Challenges

Recommended Random Forest model has recall score of 52% for positive default status. Therefore, it is suggesting that the model will miss 48% of positive instances that is actual positives. Therefore, the model is limited in predicting positive instances.

Future Uses / Additional Applications / Recommendations

Random Forest Classification model is recommended for implementation.

Implementation Plan

The model can be run on real-time basis as applications come in to raise red flags for fraud.

Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Ethical Considerations

Gender and age groups were found to be appropriately represented in the dataset and does not appear to include any bias based on these groups.

References

Carther, A., Quakenbush, C., & McKernan, S. M. (2022, March 21). The Number of Americans with Debt in Collections Fell during the Pandemic to 64 million. Urban.org. https://www.urban.org/urban-wire/number-americans-debt-collections-fell-during-pandemic-64-million

Wood, T. (n.d.). What is Random Forest? Deepai.org. https://deepai.org/machine-learning-glossary-and-terms/random-forest

10 Potential Questions

1. Did you check for overfitting?

   Yes, overfitting was checked utilizing training and test accuracies. For random forest, model overfitting was noticed as training accuracy was 1 while test accuracy was ~87%. For Naïve Bayesian model, no overfitting was noticed as both training and test accuracies were close to 74%.

2. Why is Naïve Bayes model performance lower? Can you add some insights?

   Naive Bayes assumes that all features are independent of each other given the class label. However, in this case, some of the features as seen in the heatmap are correlated. In addition, this model does not perform well in case of continuous data. These limitations have resulted in lower model performance.

Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

3. What is your interpretation of f1 score?

   For Random Forest and Naïve Bayes models f1 scores for prediction of no-default or value 0 were higher whereas they were 66% and 5% respectively for prediction of default or value 1. F1 score is a balanced score using both precision and recall and a higher f1 score indicates robustness of the model. Random Forest model is more robust in predicting default.

4. Did you investigate if income and regions are correlated based on high default cases in some regions?

   Income and regions have minimal correlation as noted in the correlation report. Northeast and central had -0.015 correlation coefficient whereas South had 0.002.

5. How can you conclusively infer from your data that females pay on time and not prone to default compared to males?

   We will have to run hypothesis test with null hypothesis that females do not pay on time compared to males together with alternate hypothesis that they pay on time and consider a significance level (0.01 or 0.05). If p value comes lower than the chosen significance level, we can reject the null hypothesis and accept alternate hypothesis.

6. Did you check your models including the features that you dropped based on the number of missing data?

   Rate of interest, Interest rate spread, upfront charges, property value, LTV, and dtir1 were dropped. Of these only Loan to Value or LTV appear to add some risk to the loan. However, it was determined that loan limit feature would influence the LTV value and therefore should be indirectly covered by it.

7. What is the train-test split ratio?

80%:20%

8. What are the business implications of model prediction?

   Business implication is improvement in predicting the potential of default and therefore, maintaining profitability of the lending institution.

9. Is your data not dated?

   Data is from 2022. Inflation is at the highest level at present and therefore, financial constraints on the borrowers should be higher than they were in 2022. Therefore, the data may be dated for current predictive models.

10. How do you tend to keep the model relevant?

    Model can be kept relevant by using updated data and validating against the number of actual default cases.

Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Appendix

Figure 1

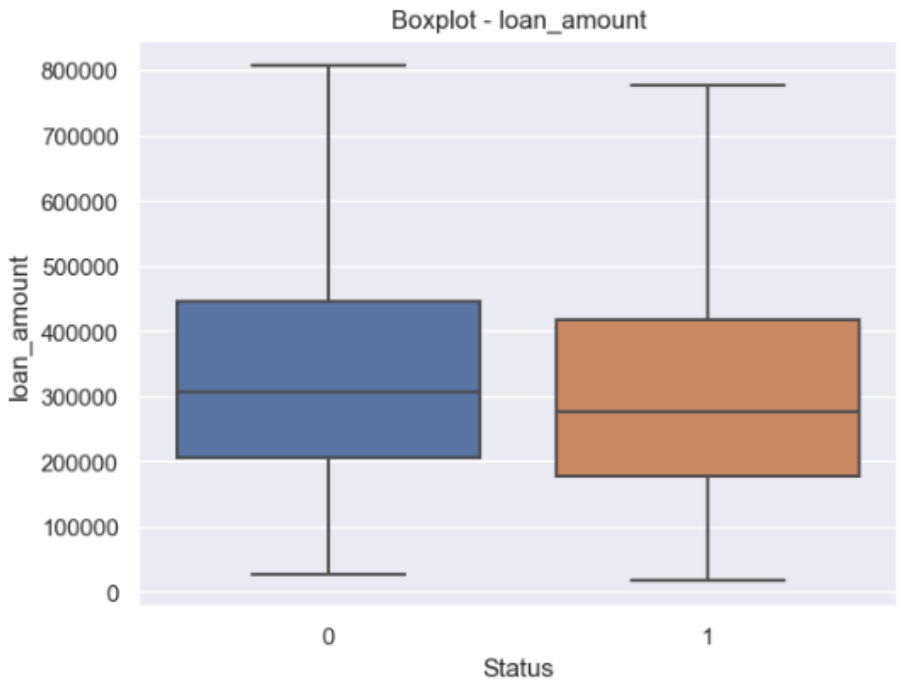
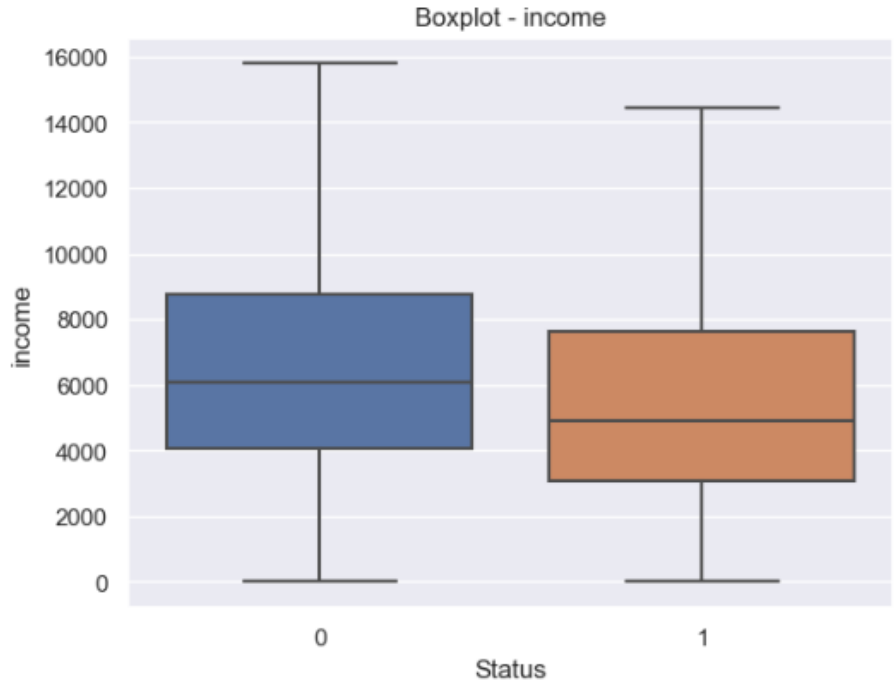


Figure 2



Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Figure 3

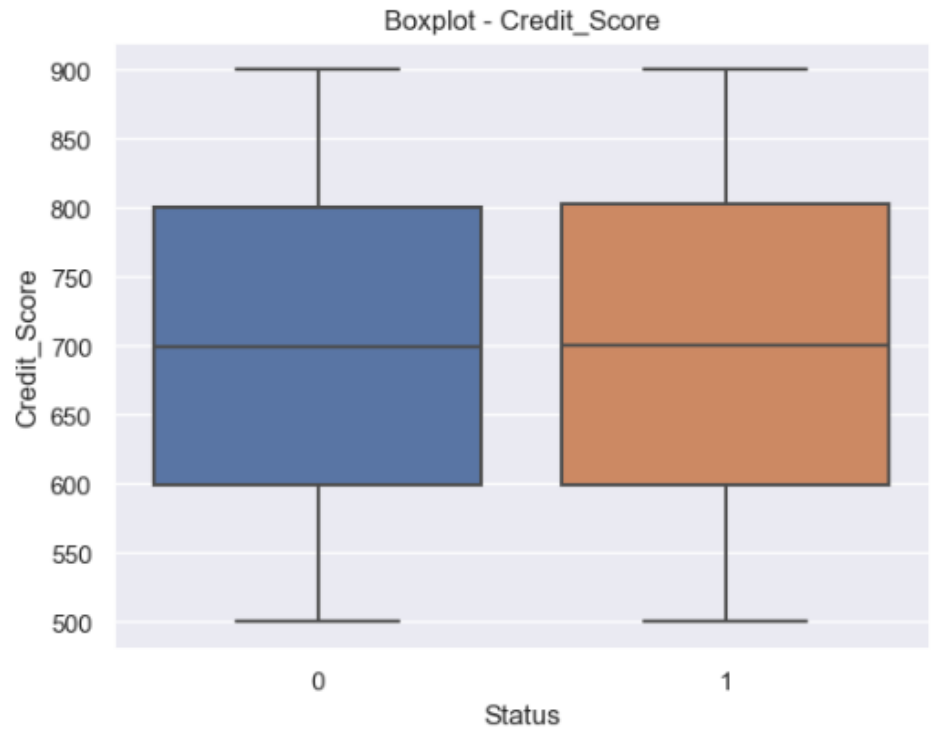
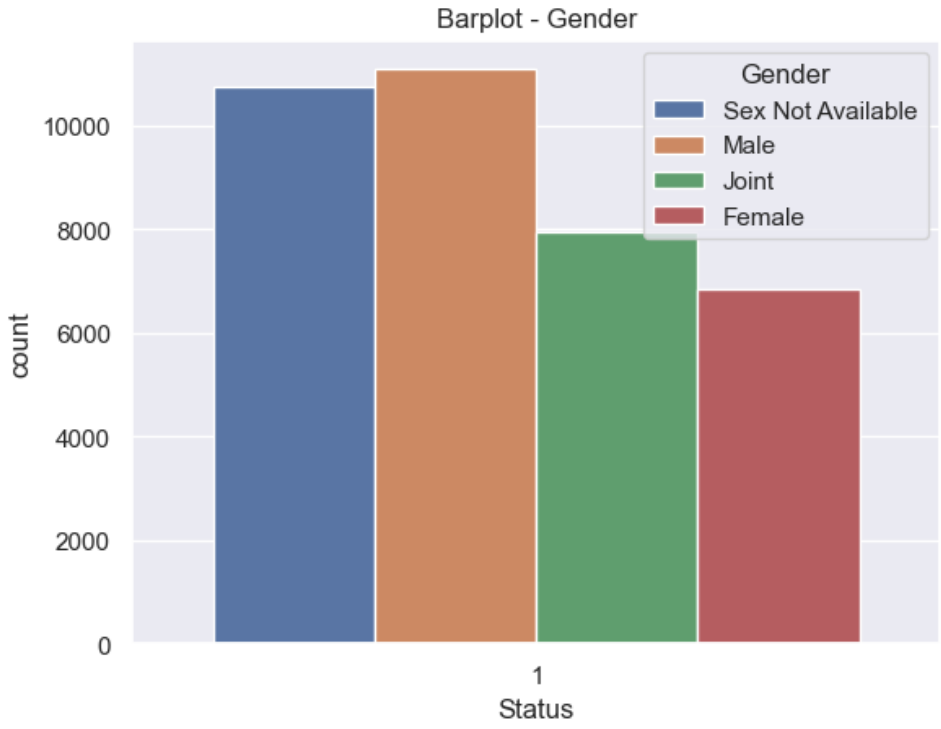


Figure 4



Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Figure 5

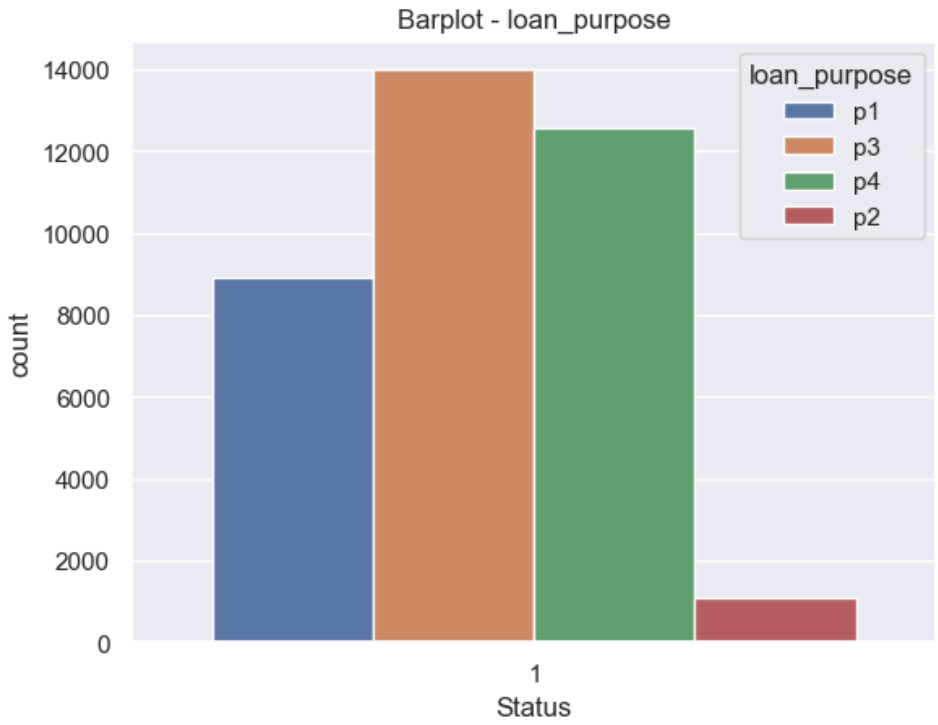
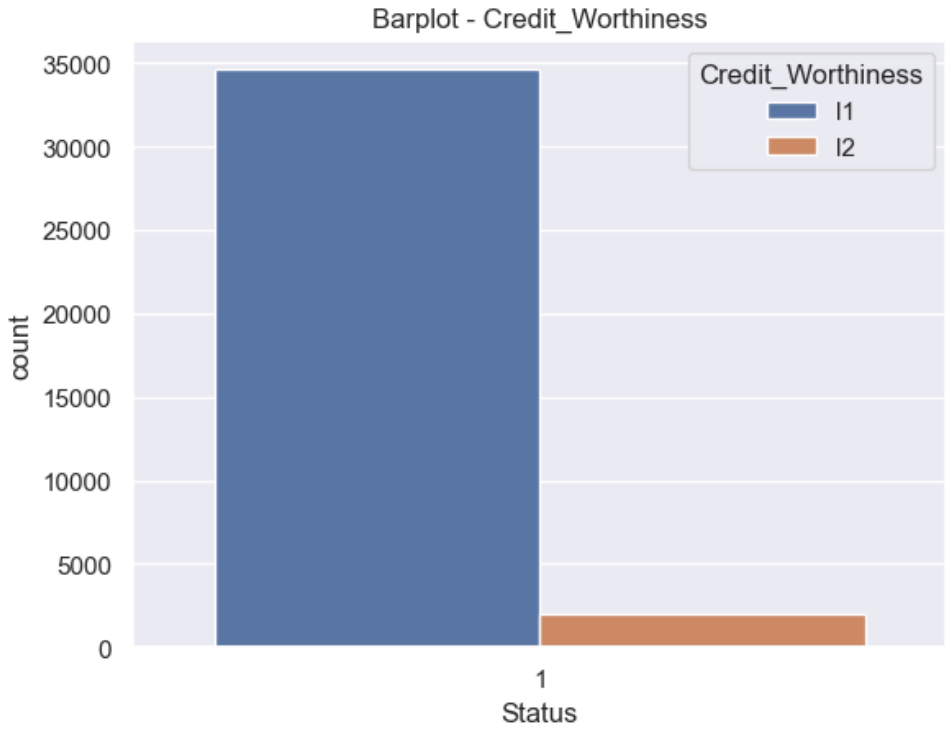


Figure 6



Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Figure 7

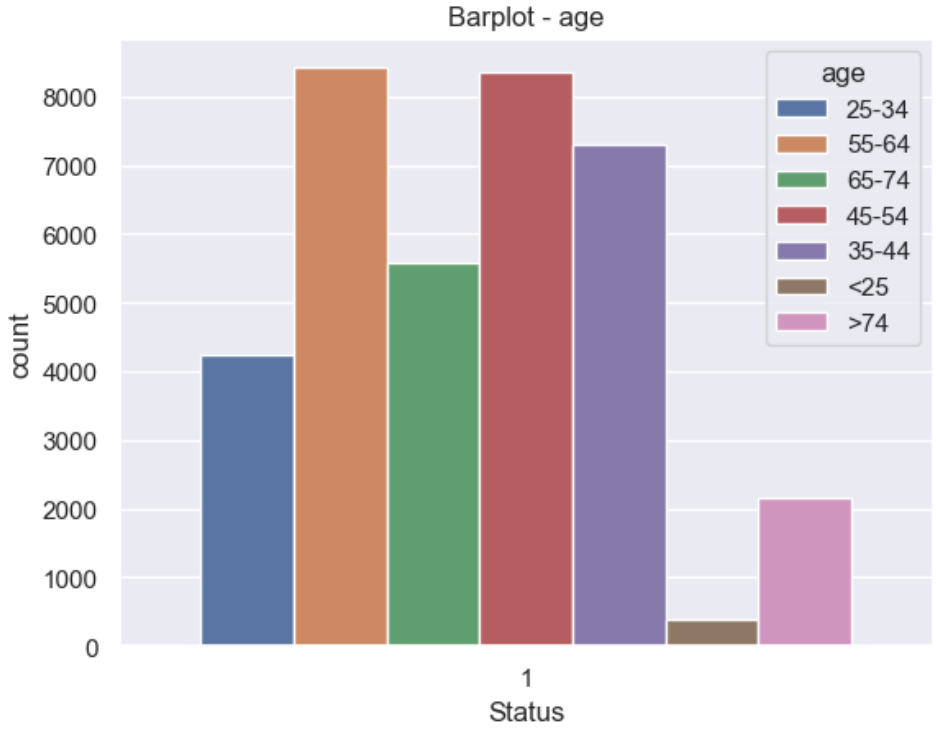
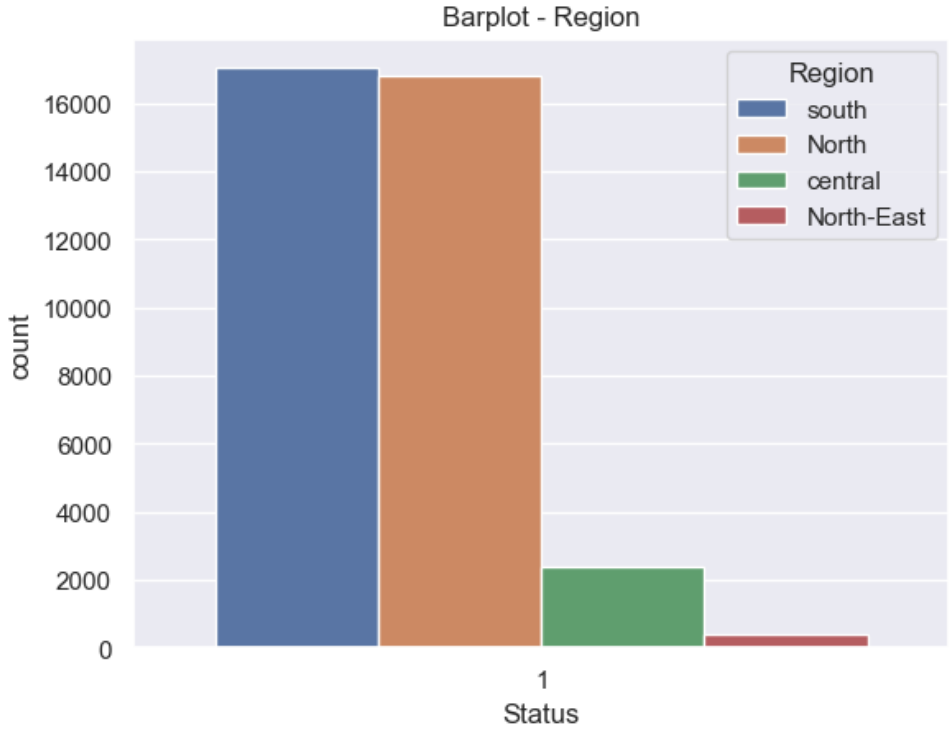


Figure 8



Shashi Bhushan
05/05/2024
DSC680 – Applied Data Science
Spring 2024

Figure 9

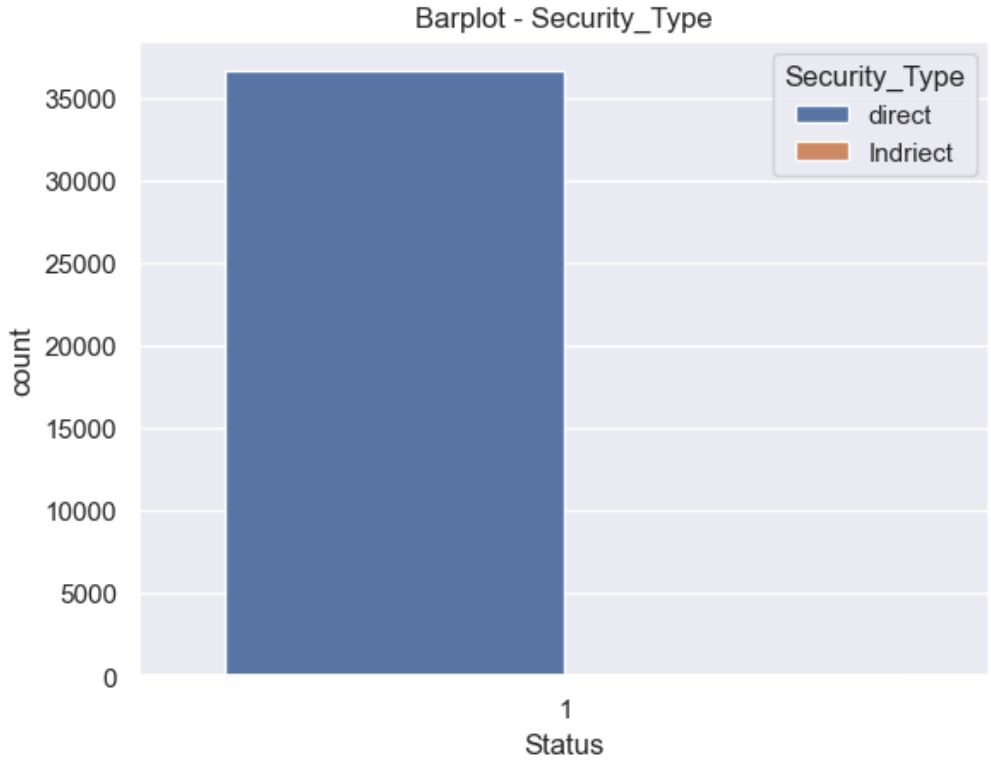


Figure 10

	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