Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

Project Milestone 3

Business Problem

Credit card fraud is an unauthorized use of the monetary instrument by someone other than the cardholder. This can impact businesses and individuals negatively financially as well as emotionally for a potentially prolonged period requiring responses by bank authorities and law enforcement agencies thereby creating additional financial burden on the society. This clearly represents a problem that requires to be solved or a solution being made where its impacts are minimized.

Background/History

The Federal Trade Commission reports that credit card frauds are a big issue which users should be aware of and protect against with about 390,000 credit card identity theft cases being reported to the agency in 2021 alone. The modus operandi employed by the fraudsters included using emails, text messages, or phone calls to get a consumer's card information and utilize the information gained to conduct unauthorized transactions. In addition, skimming frauds are also common where card information is copied at point of sales kiosks (Johnson, 2023). Therefore, we see that credit card frauds represent a challenge to the wellbeing of a society and require a lot of financial & technology investments to minimize the occurrence.

Data Explanation

Source - Credit Card Fraud Prediction (kaggle.com).

The dataset has 22 variables with fraud being the target variable with answers noted as binary (0 or 1). It has 555,719 instances in total containing a mix of data type (categorical and numerical).

Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

Features include transaction date, credit card number, merchant name, merchant category (personal care, health fitness, childcare, etc.) amount, names, gender, street address, city, zip code, location (latitude, longitude), job, and date of birth of the card holder, transaction identification number, transaction time stamp, merchant location (latitude, longitude), and the target variable Is_fraud.

These features should be sufficient to train a model for the binary classification problem of existence or not of credit card fraud.

Data Preparation

Some of the features such as sequence id, or unnamed first column, transaction date and time, transaction number, merchant name, unix time, street, customer identification number, customer first name & last name were removed from the data frame. Age column was introduced based on date of birth of the customer. After the Exploratory data Analysis was completed using visualizations, state and dob features were also dropped.

Given the number of categories available in the features category, city, and job, label encoding was selected to transform / encode categorical variables of the aforementioned categories together with gender. Dummy variables creation would have created multiple variables thereby demanding more memory usage.

Methods

Visualizations

I used the following visualizations to explain my data:

- Bar chart showing count of males and females who were subjected to fraud (Figure 1).

- Bar chart showing positive counts of fraud by state (Figure 3).
- Bar chart showing positive counts of fraud by merchant category (Figure 5).
- Age distribution (histogram for positive fraud cases (Figure 7).
- Associated pie charts for the above bar charts showing percentages proportions of categories (gender (Figure 2), state (Figure 4), merchant type(Figure 6)) subjected to fraud.

Analysis / Modeling

Our target variable is_fraud (1or 0) is binary. I built and evaluated two models namely Logistic Regression and K Nearest Neighbor. Used K as three for KN Classifier model as the model produced optimum results.

Logistic regression is helpful in the prediction of classification problems and involving continuous or discrete predictor variables. It also provides probabilities associated with new data. It also identifies variables that are effective in making predictions.

The nearest neighbor algorithm uses proximity to predict the grouping of an individual data point. Here, we have to predict the credit card fraud based on multiple variables; therefore, nearest neighbor is an appropriate algorithm to use.

Interpretation of Analysis / Model Results

Visualizations

1. Females are more prone to experience credit card fraud than males in the United States.
2. States such as Texas, New York, and Pennsylvania had more than 100 cases of fraud whereas Hawaii, Colorado, Wyoming, and DC had minimum cases of fraud.
3. Shopping on websites or internet-based shopping and groceries at point of sales had the maximum occurrences of credit card fraud.

4. People in the 30-70 age group appear to experience credit card fraud more than the remaining categories.

Modeling / Interpretation

Three models produced high accuracy percentages with Logistic Regression and KNN producing 99.58%, and 99.70%, respectively. However, upon generation of the classification reports for the two models, it was observed that though the model precision, recall, and f-1 scores are very high for predicting "No fraud or 0;" however, these scores are not good for predicting "fraud or 1." KNN Classifier has a decent 62% as precision score for predicting credit card fraud.

Conclusion

Precision is the proportion of every observation predicted to be positive that is actually positive. KNN model had higher accuracy together with highest precision scores for predicting credit card fraud between the two models created (refer to the classification report Figure 8 below). Therefore, I recommend the use of KNN classifier model for predicting credit card fraud.

Assumptions

I assumed that features such as transaction date and time, transaction number, merchant name, unix time, street, customer identification number, customer first name & last name will have no bearing on the model.

Limitations and Challenges

Data is highly skewed in favor of legitimate transactions making it hard to detect cases of fraud. Therefore, balancing data is challenging for such instances. The recommended KNN model predicts

cases of no fraud with a precision score of 1 whereas cases of fraud with only 0.62 precision score. Therefore, out of every 100 case which is predicted to be a fraud case only 62 will be the actual fraud cases. There will be 38 false negatives. The model works but with limitations.

Future Uses/Additional Applications/Recommendations

KNN model is recommended for implementation.

Implementation Plan

The model can be run on real-time basis as background application to raise red flags for fraud.

Ethical Assessment

I checked for gender representation in the data set and found them to be meeting the requirements for normalcy (>30). Central Limit theorem indicates use of minimum sample size of 30 for a dataset to be assumed for normalcy (Ganti, 2023).

References

Ganti, A. (2023, March 10). Central Limit Theorem (CLT): Definition and Key Characteristics. Retrieved from Investopedia: https://www.investopedia.com/terms/c/central_limit_theorem.asp#:~:text=A%20sample%20size%20of%2030,representative%20of%20your%20population%20set.

Johnson, Holly D. “Biggest Credit Card Scams to Watch out for in 2022.” Bankrate, 23 Jan. 2023, www.bankrate.com/finance/credit-cards/biggest-credit-card-scams/.

B. (2020, August 29). Credit Card Fraud Detection. Medium. https://medium.com/total-data-science/credit-card-fraud-detection-data-science-projects-fc64216849b8.

Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

10 Potential Questions with Answers

1. Did you check for overfitting?

   Yes overfitting was checked utilizing training and test accuracies. For logistic regression, training accuracy was slightly more than test accuracy, therefore the model showed minor overfitting.

   Similarly, KNN model was observed to be slightly overfitting as well.

2. How did you overcome imbalanced data problem?

   I used the metric precision that accounts for both true positives and false negatives. Relying only on accuracy would not have been correct for an imbalanced model.

3. What is your interpretation of recall score?

   Recall quantifies correct positive predictions made from all actual positive instances. For logistic regression, for fraud cases prediction (1), it was 0 whereas it was 46% by KNN model. Therefore, KNN model performed better when considering recall values.

4. Did you check whether fraud cases are more the states/cities where population is more? In other way, number of cases are aligned with population.

   Only 32% correlation was observed.

```
# find correlation
corr = np.corrcoef(df2_state['count'],df2_state['population'])
corr
```

```
array([[1.        , 0.32183026],
       [0.32183026, 1.        ]])
```

Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

5. How can you conclusively infer from your data that females are more prone to be subjected to fraud just based on the proportion you have in Figure 2?

   Correct, we cannot conclusively claim that. Hypothesis test with null hypothesis that male proportion is more than the female proportion for the entire population. We get the p vale of ~0.57 and therefore can not reject the null hypothesis and conclusively claim that female proportion subject to fraud will be applicable for entire population.

```
from statsmodels.stats.proportion import proportions_ztest
count = np.array([981,1164])
nobs = np.array([250833,304886])

stat, pval = proportions_ztest(count, nobs)
print(f"P-value: {pval:.3f}")
```

```
P-value: 0.577
```

6. Is your use of latitude / longitude data relevant to the model? Did you check model performance excluding these features?

   Latitude / longitude data is not relevant to the model. Actually, the precision and f1-score improves for prediction of fraud (1) after these features are excluded.

7. What is the train-test split ratio?

   80%:20%

8. What are the business implications of model prediction?

   Business implication is improvement in detecting the fraud and cost savings to the company as well as the reputation of the business entity and the financial institution.

9. Is your data not dated?

Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

Data is from 2020 . Though the patterns may have changed since the time data was collected.

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 fraud cases.

Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

Appendix®

Figure 1

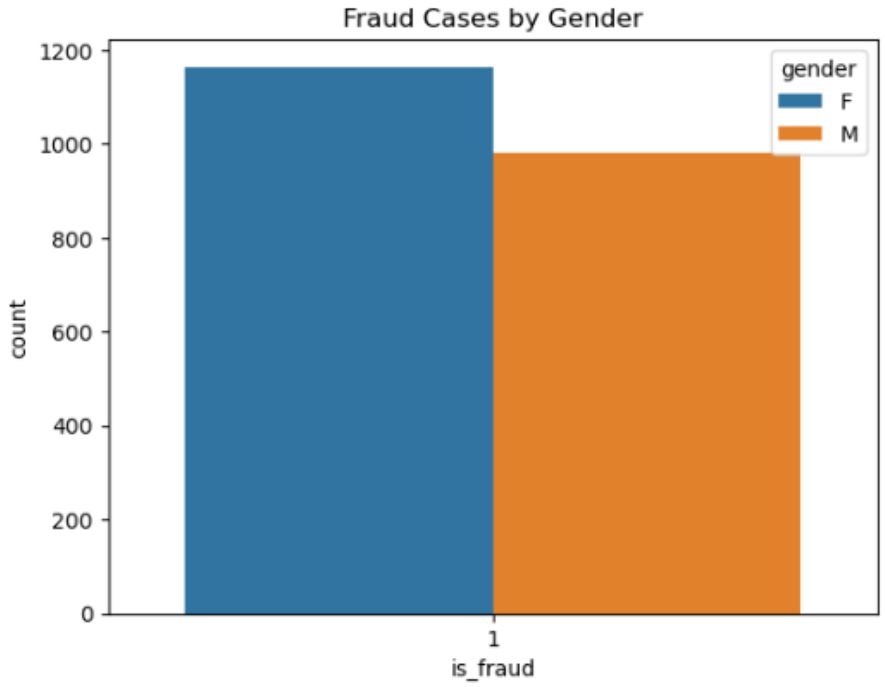
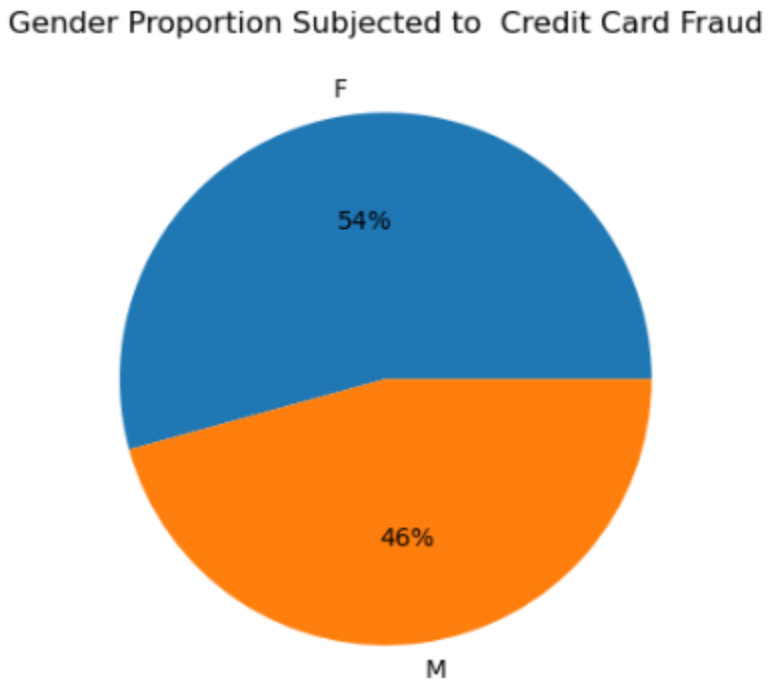


Figure 2



Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

Figure 3

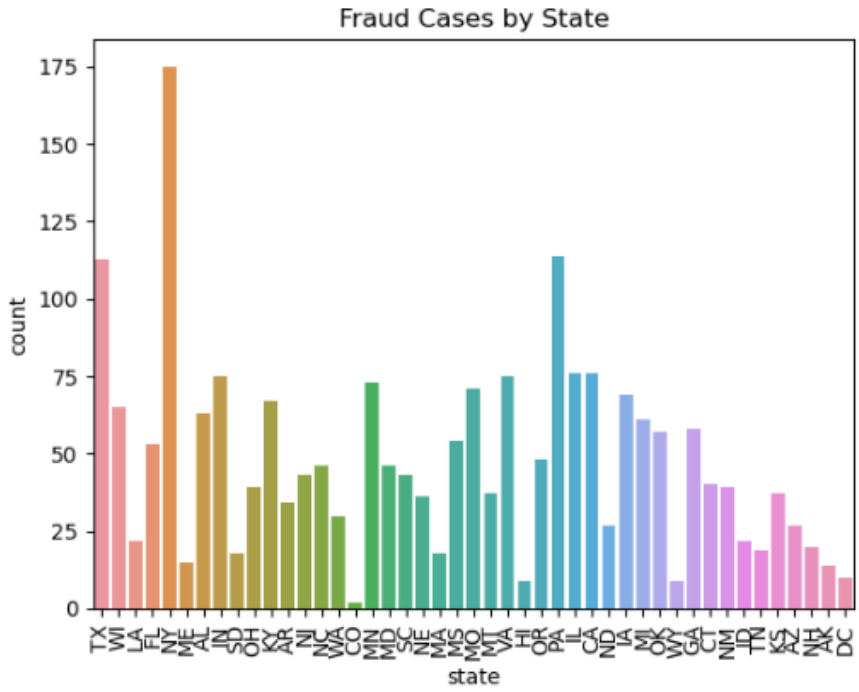
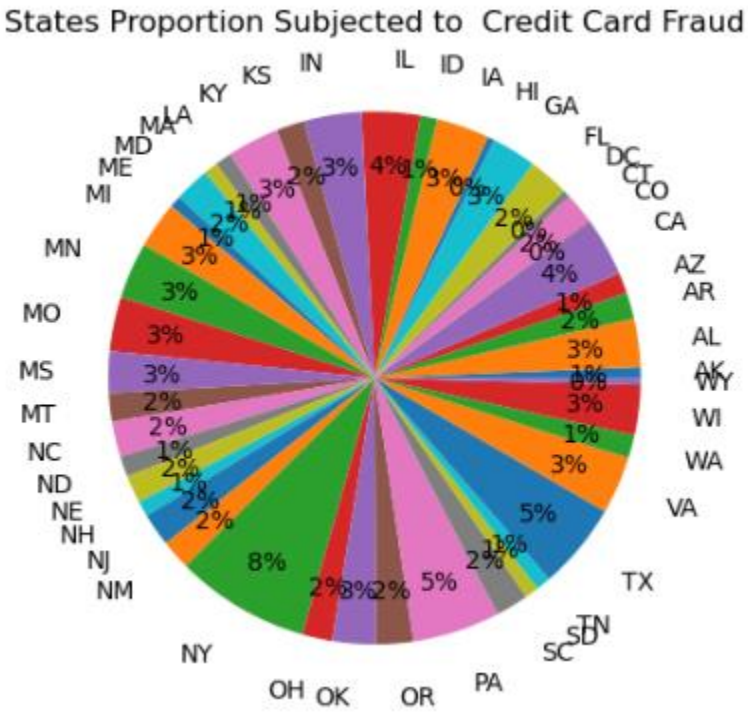


Figure 4



Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

Figure 5

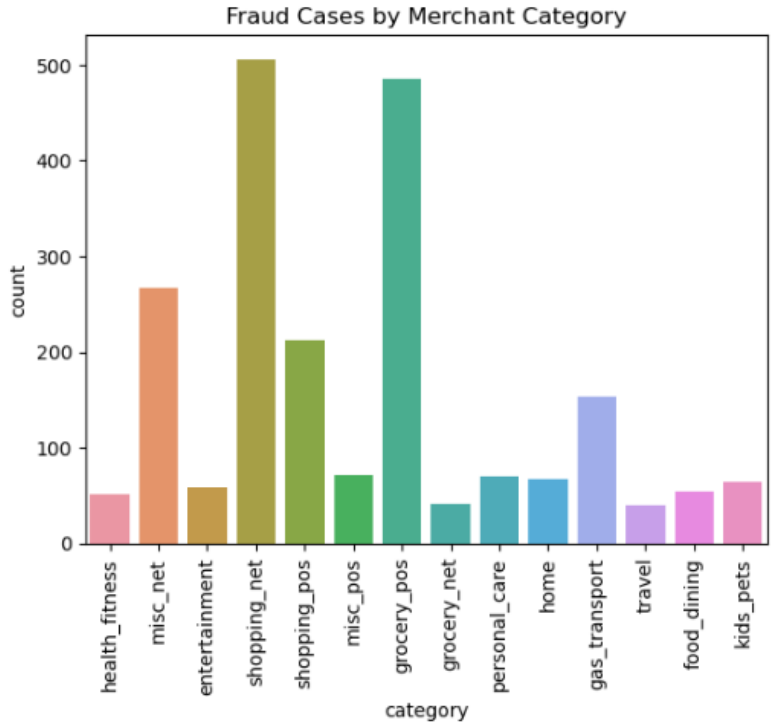
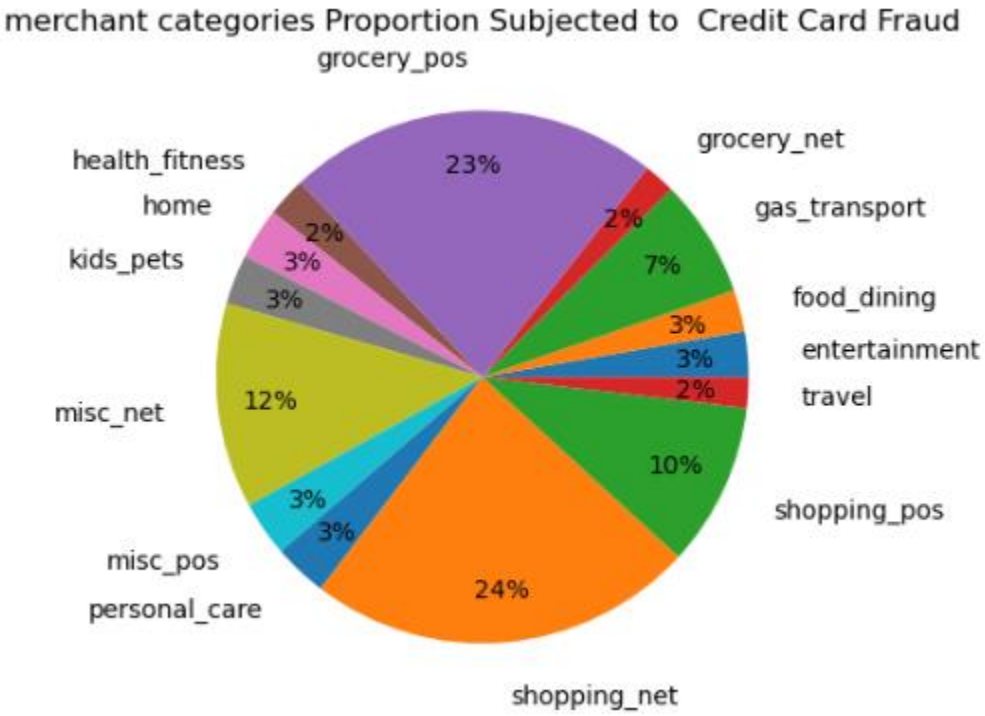


Figure 6



Shashi Bhushan
4/7/2024
DSC680 – Applied Data Science
Spring 2024

Figure 7

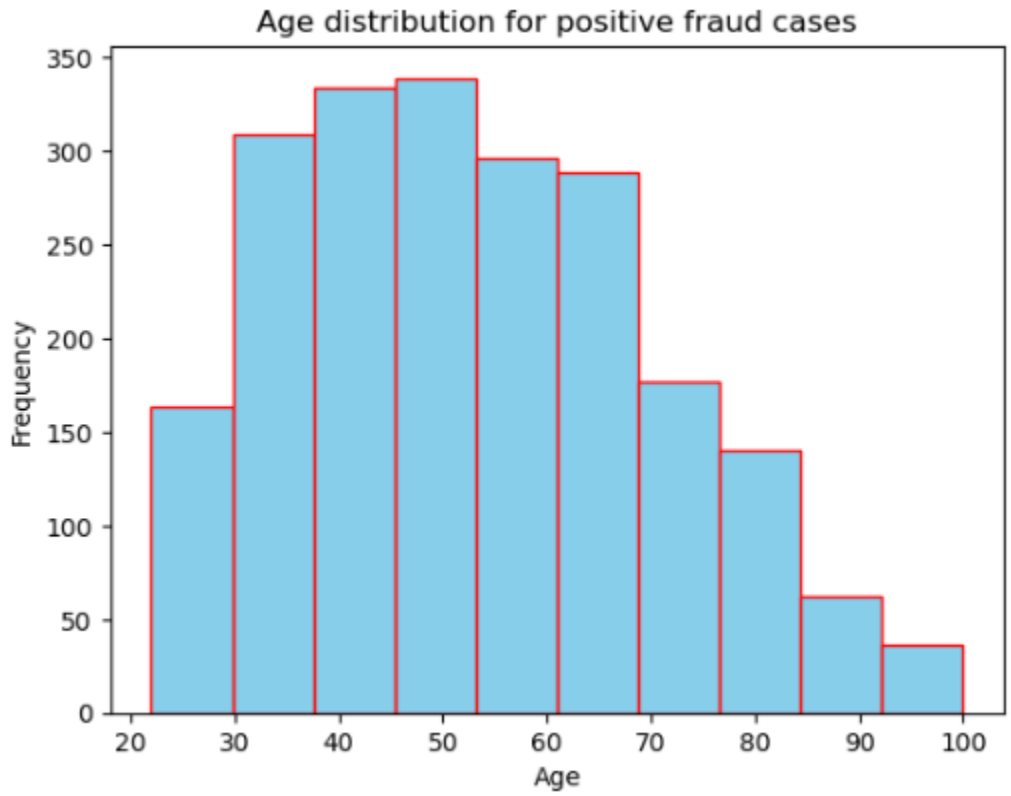


Figure 8

	precision	recall	f1-score	support
0	1.00	1.00	1.00	110734
1	0.62	0.54	0.58	410
accuracy			1.00	111144
macro avg	0.81	0.77	0.79	111144
weighted avg	1.00	1.00	1.00	111144