William Barker

DSC680

Project 3

Project 3 - Final

Business Problem

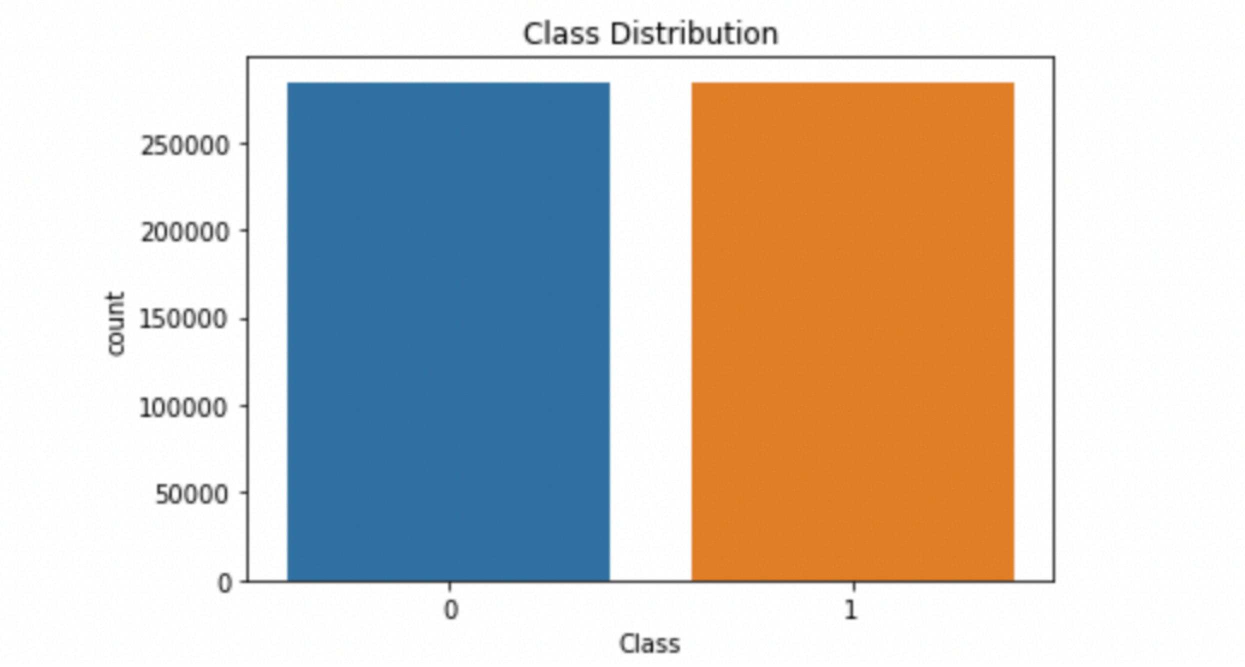
Credit card fraud is major problem and can cause a great amount of stress on those who fall victim to it. It affects businesses just as much as consumers however, if not more. Fraud.net says, “While individual victims are usually able to recover their stolen money via reimbursement and chargebacks, the merchants on the other end of the fraudulent purchases are often left hanging.” This project would allow credit card companies to hopefully detect fraud as quickly as possible so cards can be cancelled before more purchases are made.

Background/History

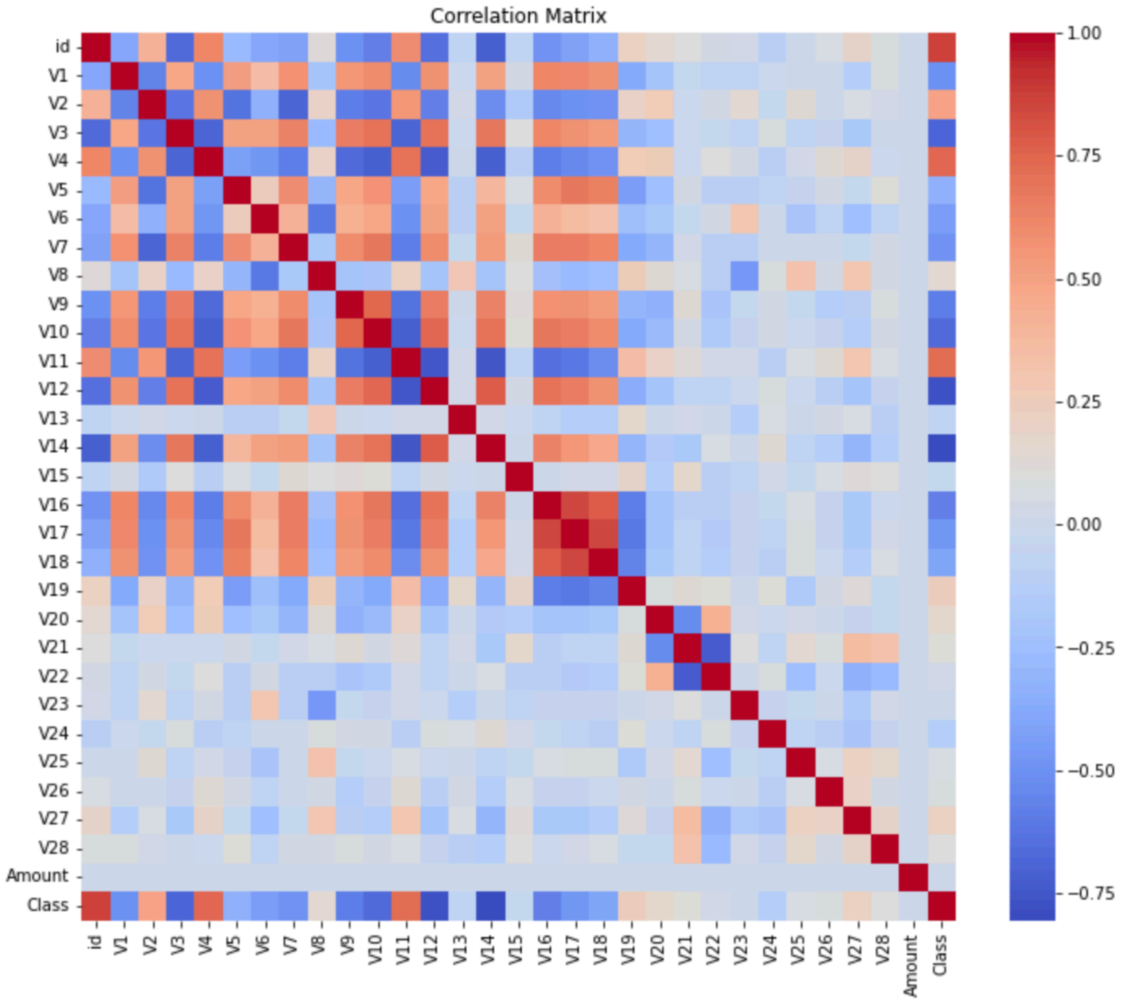
“The very first instance of credit card fraud is reported to have occurred in 1899 when the early precursors of credit cards were just emerging. The story goes: a farmer threw out his credit card and a crook allegedly picked it up and used it to buy $25 worth of train travel. In the 1960s, credit card fraud was based on forging cards with the help of expensive machines and taking advantage of the banks’ new, unregulated system of credit by fraudulently acquiring cards and racking up purchases. In 1970, Life magazine reported on the novel case of the Orlick family, who ended up on the hook for thousands — despite their credit limit of $400 — after someone physically stole their brand-new, bank-issued credit card. In 1983, MasterCard and Visa introduced complex backgrounds, logo holograms, and other design elements to their cards in order to make the physical cards more difficult to reproduce. In the mid ‘90s, internet phishing was born. AOHell, a malicious program created to defraud those with AOL dial-up accounts, was one of the first phishing scams that was widely used to steal credit card information. As credit cards began to incorporate magnetic stripes in the 1990s, fraud evolved to match. Criminals began changing the magnetic coding in order to commit fraud, a trick that remained popular throughout the 2000s. Also in the 2000s, “carding websites” — online marketplaces where people bought and sold stolen credit card information — exploded in popularity. In the 2000s, in response to increasing credit card fraud, EMV technology was born. EMV cards use an electronic “chip” and a four-digit Personal Identification Number (PIN) to securely authorize transactions. EMV became the new standard in credit card technology by the mid-2010s. Over the last 15 years, millions of sets of payment card information have been stolen from Adobe Systems, Target Corporation, Neiman-Marcus, Home Depot, and more, in high-stakes cases that have affected consumers around the globe. Phishing also remains common to this day, with notable examples affecting banks, the United States’ defense suppliers, airlines, Sony, and more. In recent years, the world has seen as dramatic increase in CNP (Card Not Present) fraud, which occurs when stolen credit card information is used to make a transaction from a distance.”

Data Explanation

The dataset I'm using for this project is one from Kaggle.com titled creditcard_2023. It's one of the largest datasets I've ever worked with, having 568,631 rows and 31 columns. 28 of the columns are anonymized features representing various transaction attributes like time and location. The others are "id", "Amount" and "Class". "Amount" being the transaction amount and "Class" being a binary label indicating whether the transaction is fraudulent (1) or not (0). I used "describe" to get summary statistics of my dataset, then checked for any missing values, and didn't find any. I checked the distribution of the "Class" column and found a perfectly even amount of fraudulent and non-fraudulent transactions.



I then visualized a correlation matrix to check for possible correlations between features.



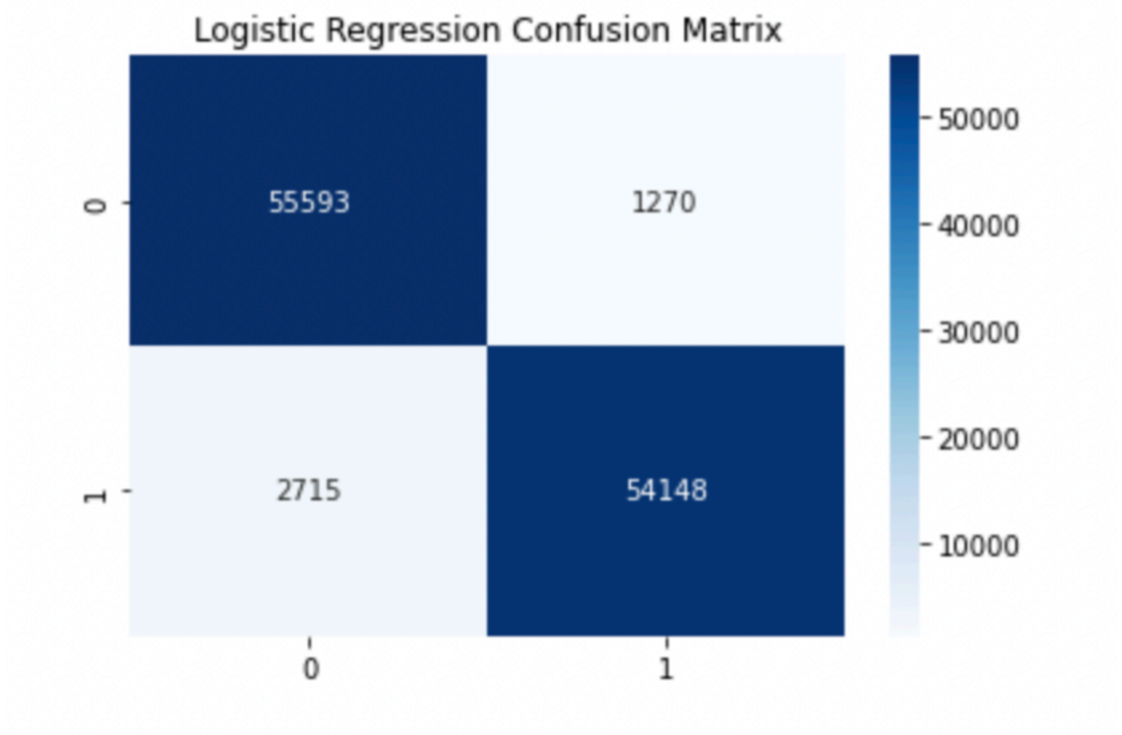
Finally, I dropped the “id” column since it serves no purpose for me and split my data into a training and test set, and scaled the features.

Methods

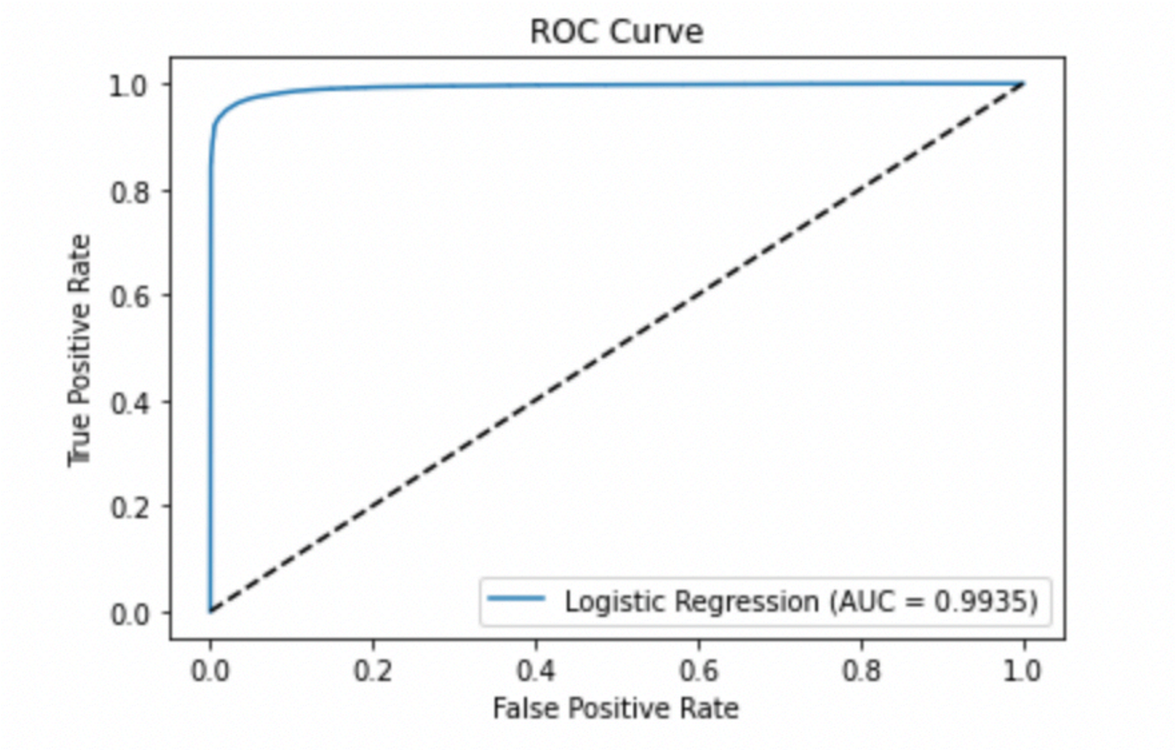
For this project, I implemented the simple yet powerful linear regression model to try and predict whether a transaction is fraudulent or not. I also branched out and tried something new for me, an isolation forest model which is used to detect anomalies.

Analysis

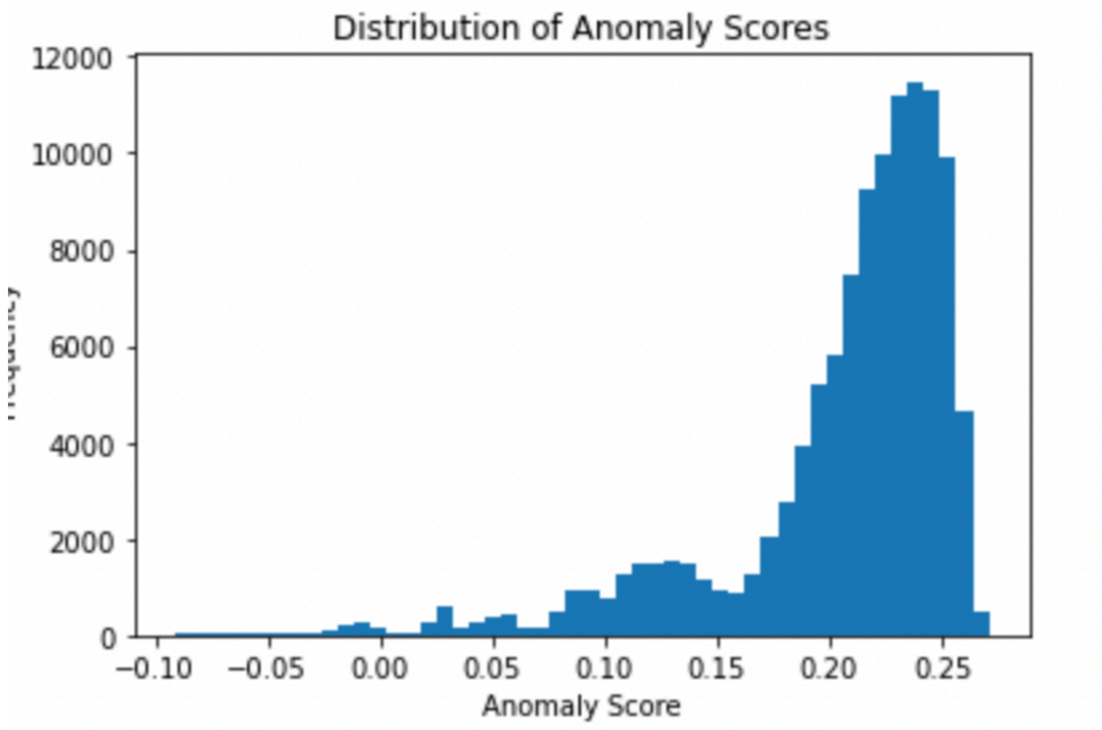
I started with training my linear regression model. I then made predictions with the scaled data and used a classification report to evaluate the results. I got a precision score of 0.95 for non-fraudulent transactions and 0.98 for fraudulent ones, a recall score of 0.98 for non-fraudulent transactions and 0.95 for fraudulent ones, and an f1-score of 0.97 for non-fraudulent transactions and 0.96 for fraudulent ones. I also got an accuracy score of 0.96. Precision is the ratio of correctly predicted positive observations to the total predicted positives. High precision indicates that the model doesn't classify many non-fraudulent transactions as fraudulent. In other words, it tells you how accurate the positive predictions are. Recall is the ratio of correctly predicted positive observations to all observations in the actual class. High recall indicates that the model captures most of the actual fraudulent transactions. It tells you how well the model is at identifying frauds. The F1-Score is the weighted average of Precision and Recall. A high F1-Score means that the model has a good balance between precision and recall, making it more reliable in the context of fraud detection where both metrics are important. Accuracy is the ratio of correctly predicted observations to the total observations. Accuracy gives an overall performance measure. My scores with my linear regression model are incredible! With an f1-score of 0.97 and 0.96 and an accuracy score of 0.96, my model is almost always correct in its predictions. I built a confusion matrix which helps to visualize just how accurate the model was.



I also got an ROC-AUC score of 0.9935, which is very high and indicates my model is very effective!



Next I initialized and trained my isolation forest model. I predicted the anomalies of my data and then converted the predictions to match the binary classification system of my dataset. I once again used a classification report to evaluate my model. I got a precision score of 0.50 for non-fraudulent transactions and 0.97 for fraudulent ones, a recall score of 1.00 for non-fraudulent transactions and 0.02 for fraudulent ones, and an f1-score of 0.67 for non-fraudulent transactions and 0.04 for fraudulent ones. I also got an accuracy score of 0.51. The only score that is good is the precision score for fraudulent transactions, which makes sense because the isolation forest model is an anomaly detection algorithm. This score means that the majority of the time my model detected an anomaly (fraudulent activity), it was correct! I also got the anomaly scores for my test set and visualized the distribution.



Conclusion

In conclusion, I believe both of my models to be extremely effective at what they do! My linear regression is able to very accurately make predictions on the data and my isolation forest model did a good job and accurately detecting fraud.

Limitations

I only trained two models for this project, but if I had tried some more powerful models I may have possibly gotten even more accurate predictions.

Challenges

This was my first experience with an isolation forest model, and I wasn’t 100% certain the best way to evaluate its accuracy.

Future Uses

These models could be used by credit card companies to accurately predict and/or flag fraudulent credit card transactions!

Ethical Assessment

Being able to look at someones transaction data could be considered an invasion of privacy, since the data could be used to track someones location at any given time, what they are buying etc. Luckily, the dataset I am using completely anonymizes all transaction attributes and does not include anyone’s name, so I think ethically I’m ok.

References

https://fraud.net/n/how-does-credit-card-fraud-affect-businesses/

https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023?resource=download

https://www.ellipse.la/post/a-brief-history-of-credit-card-fraud#:~:text=The%20very%20first%20instance%20of,%2425%20worth%20of%20train%20travel.

10 Questions

1. What motivated you to choose credit card fraud detection as your project?
2. How did you preprocess the data before building the models?
3. Why did you choose Isolation Forest and Linear Regression for this task?
4. How did you evaluate the performance of your models?
5. What challenges did you face during the project, and how did you overcome them?
6. How did you handle the class imbalance in your dataset?
7. Why is it important to balance precision and recall in fraud detection?
8. What are the limitations of the models you used?
9. How could you improve the performance of your fraud detection system?
10. How do you see this project being applied in a real-world scenario?

Answers

1. Credit card fraud is a significant problem that affects both consumers and businesses. I chose this project to apply machine learning techniques in a

real-world scenario that has a direct impact on people's lives and businesses.

2. I got a description of my dataset, checked for any missing or null values, checked the distribution of the "class" feature, built a correlation matrix, dropped the "id" column, separated my features and target variable and split the data into a training and test set.
3. I chose Isolation Forest because it is well-suited for anomaly detection, which is a key aspect of fraud detection. I also used a linear regression to try and predict whether a transaction is fraudulent or not because it is simple yet powerful.
4. I used several metrics, including precision, recall, F1-score, and accuracy. For the Isolation Forest, I also examined the anomaly scores and used a confusion matrix to better understand the model's strengths and weaknesses.
5. This was my first experience with an isolation forest model, and I wasn't 100% certain the best way to evaluate its accuracy.
6. I got lucky, my dataset was perfectly balanced between fraudulent and non-fraudulent transactions, making it much easier to work with.
7. In fraud detection, precision and recall represent two important trade-offs. High precision ensures that most transactions identified as fraudulent are indeed frauds, reducing the cost of false positives. High recall ensures that most actual frauds are detected, reducing the risk of missed frauds.

Balancing these two metrics is crucial because focusing too much on one can lead to an unacceptable increase in the other, which makes the f1-score such an important metric to get a good score with.

8. One limitation of the Isolation Forest is that it assumes anomalies are sparse and distinct, which might not always be the case in complex fraud scenarios. Linear Regression is not typically used for classification tasks and may not perform well in highly non-linear or complex datasets like this one.
9. I could combine the Isolation Forest with other models like Random Forest or Gradient Boosting to create an ensemble method.
10. By integrating this model with a live transaction processing system, it could help in identifying potentially fraudulent transactions in real-time. With further refinement and testing, it could reduce the incidence of fraud and save businesses and consumers from significant financial losses.