Can We Predict Whether New Credit Card Transactions Are Fraudulent Or Not?

Ilaria Alessandrelli, Joey Chan, Juliana Novaes, Nicol Siccardi and Rumaanah Ellahi

Introduction：Why Credit Card Fraud?

Credit card fraud is as prevalent as ever, with online transactions becoming more common.

The aim of our project is to be able to detect whether new transactions are fraudulent or not through the use of different machine learning techniques, as well as to identify patterns in credit card fraud using data visualisations.

The project uses a dataset containing 23 features and other external data was also added, such as information on COVID restrictions, crime rates and weather .

Introduction: Questions We Hope To Answer

- How to predict if new transactions are fraudulent or not? - *Focus Question*
- Do areas of higher crime rate commit more Credit Card Fraud?
- What is the relationship between crime rates and fraud rates?
- Did the effects of COVID restrictions result in more Credit Card Fraud than different periods of time throughout the year?

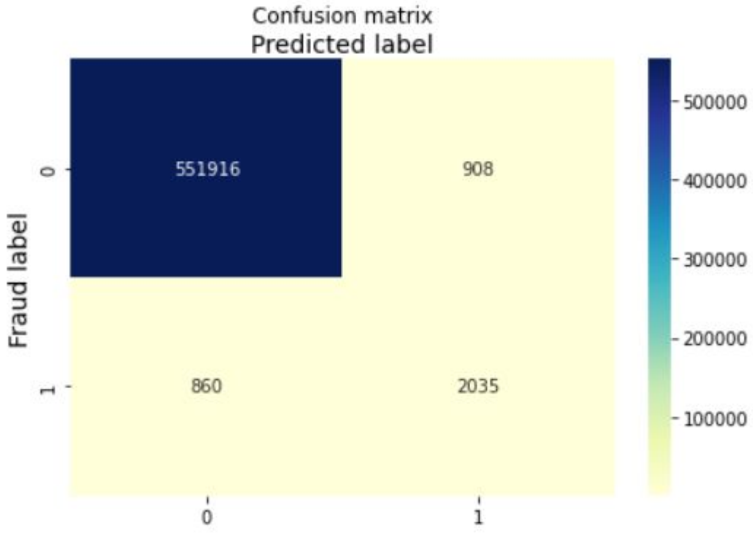
Introduction: How Can Our Analysis Help?

Credit Card Fraud analysis would be helpful to several industries, the target audience of the project would be Banks and Insurance Companies, as we believe they would benefit the most from being able to detect fraudulent activities.

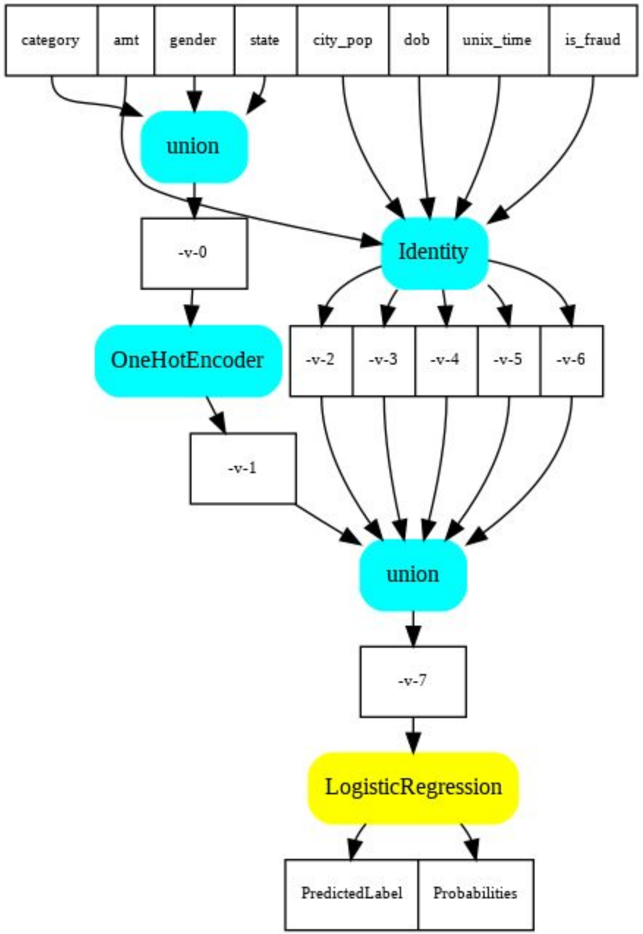
Banks could possibly use the findings to prevent the transactions from occurring and insurance companies to better protect their customers from fraud and perform better risk analysis.

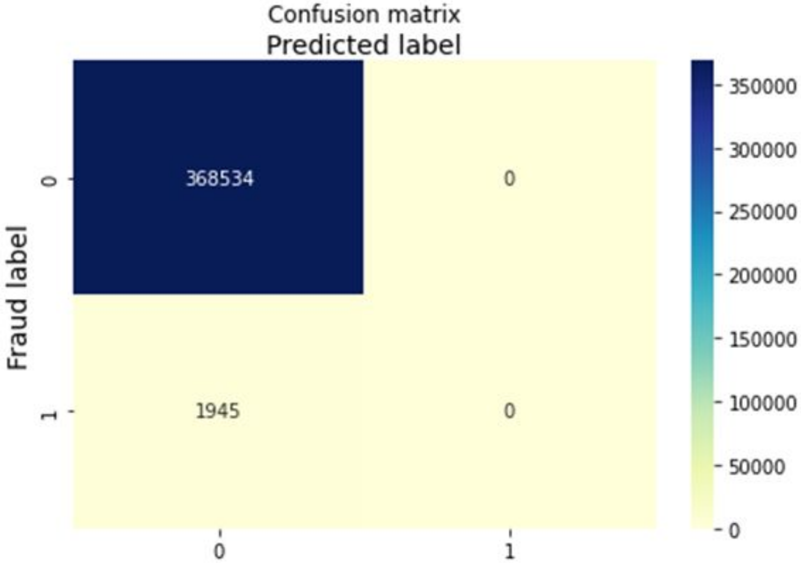
Decision tree model

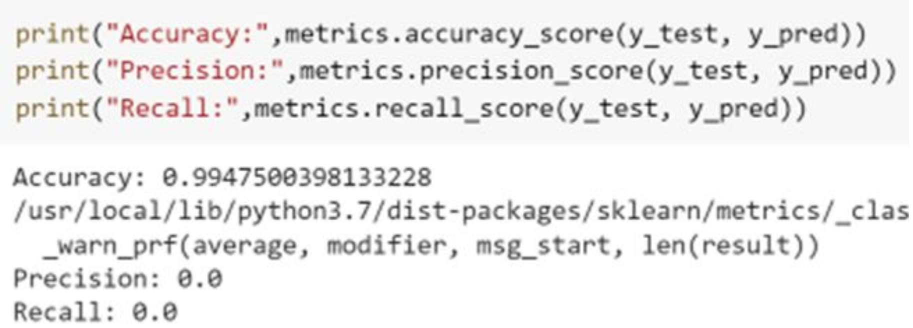
A model that builds itself as a pathway to a decision. It includes conditional 'control' statements to classify data, starting at a single point (or 'node') which then branches (or 'splits') in two or more directions. Each branch offers different possible outcomes, incorporating a variety of decisions and chance events until a final outcome is achieved.



Logistic Regression Model







```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```

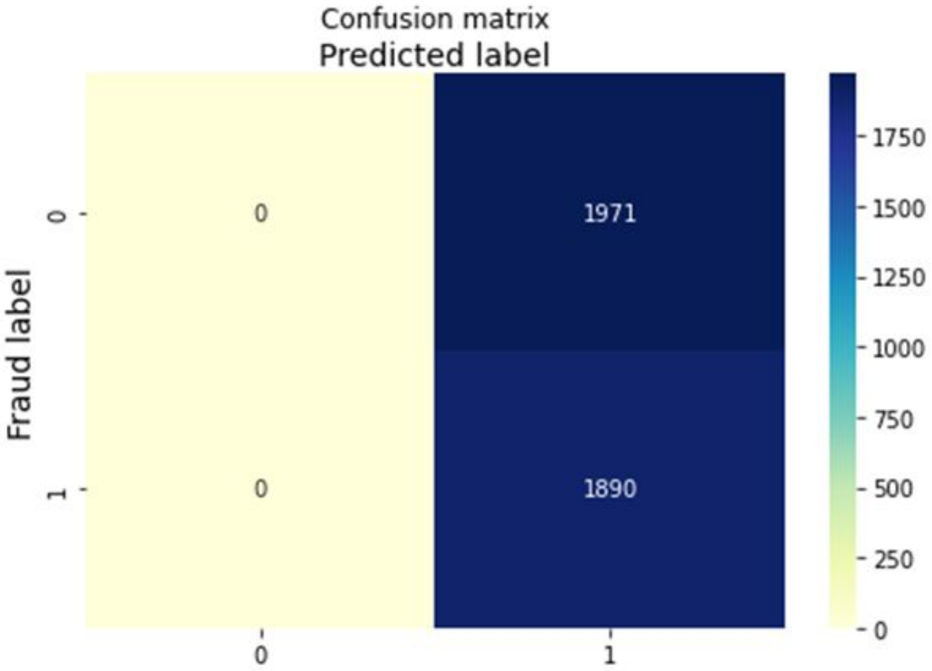
```
Accuracy: 0.9947500398133228
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_clas
  _warn_prf(average, modifier, msg_start, len(result))
Precision: 0.0
Recall: 0.0
```

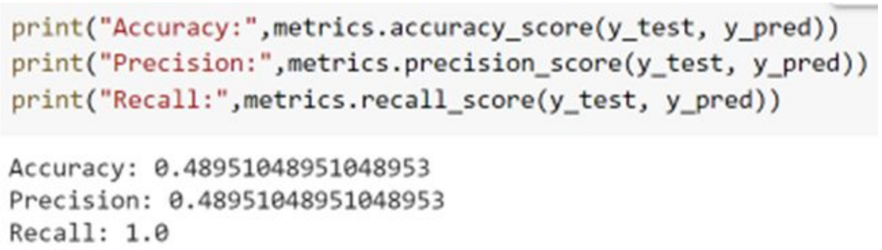
Employing Resampling to prevent Overfitting

These techniques were applied to Logistic Regression algorithms to deal with the unbalanced nature of the fraud dataset:

1. UPSAMPLING: It artificially inflates the instances of fraud present in the database by duplicating data. The resulting model was prone to overfitting.

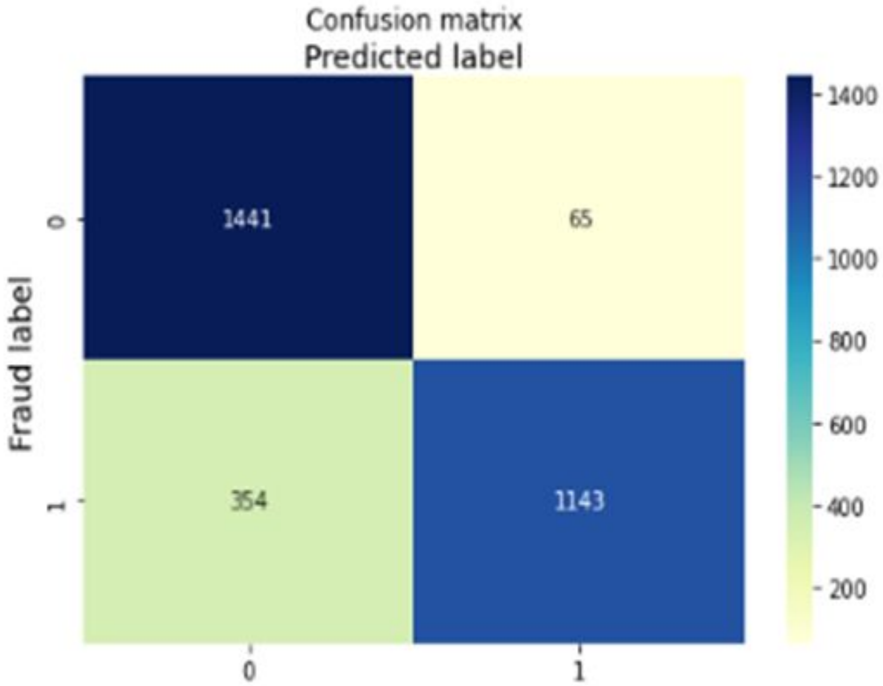
2. DOWNSAMPLING:It reduces the number of genuine transactions to match the number of fraud instances. This initially showed very promising metrics, with high accuracy, recall and precision scores. However, selecting different cuts of the reduced pool of genuine transactions produced inconsistent results.

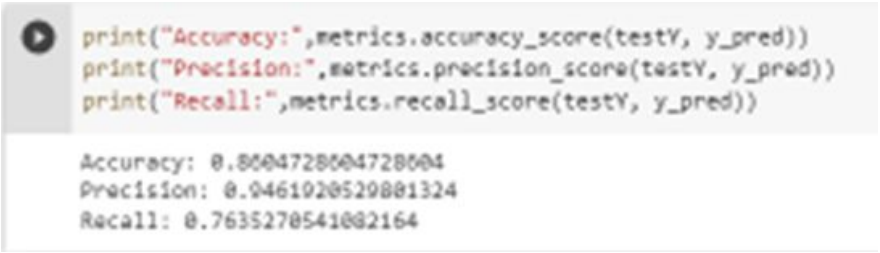




```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.48951048951048953
Precision: 0.48951048951048953
Recall: 1.0
```





```
print("Accuracy:",metrics.accuracy_score(testY, y_pred))
print("Precision:",metrics.precision_score(testY, y_pred))
print("Recall:",metrics.recall_score(testY, y_pred))
```

```
Accuracy: 0.8604728604728604
Precision: 0.9461920529801324
Recall: 0.7635270541082164
```

Employing Cross-Validation to prevent Overfitting

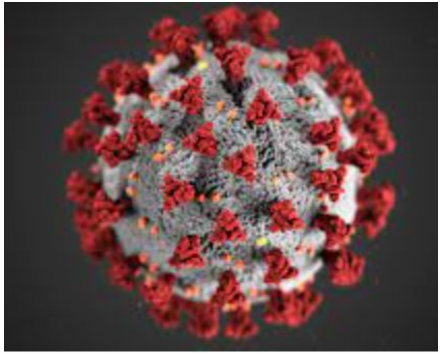
1. Tests to prevent logistic regression models from overfitting were performed with the following cross-validation techniques:
   a. KFold
   b. StratifiedKFold
   c. RepeatedStratifiedKFold

   All tests were unsuccessful in preventing overfitting.

2. Finally, we employed cross-validation within a cost-sensitive machine learning framework. This however, continued to yield a logistic regression model prone to overfitting

Model Results 1 - to be completed by Juliana/Ilaria

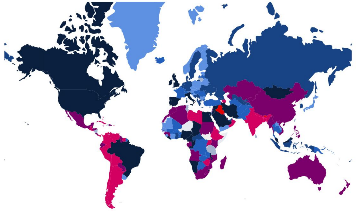
Model Results 2 - to be completed by Juliana/Ilaria



Data Gathering - APIs:

Nicol & Joey

- **Oxford COVID-19 Government Response Tracker API**
  - Did the effects of COVID restrictions result in more Credit Card Fraud than different periods of time throughout the year?
- **Federal Bureau of Investigation - Crime Data Explorer API**
  - Do areas with higher crime rate commit more Credit Card frauds?
  - What are crime rates in comparison to the Credit Card Fraud rates?



Oxford COVID-19 Government Response Tracker API (Nicol)

- Track and compare government responses to the coronavirus.
- >180 countries
- Data collected daily since 1st January 2020
- Stringency Index
- Nested dictionary

BLAVATNIK
SCHOOL OF
GOVERNMENT

UNIVERSITY OF
OXFORD

Federal Bureau of Investigation - Crime Data Explorer API (Joey)

- Having only Nationally accessible data
- Minimal documentation
- Understanding the various APIs available to delve into them
- The extraction of data from a JSON file to a CSV file

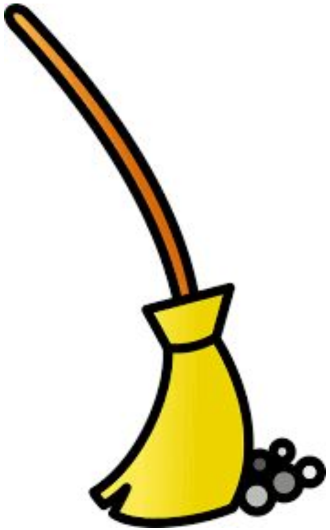
Data visualisations & analysis - APIs:

Joey & Nicol

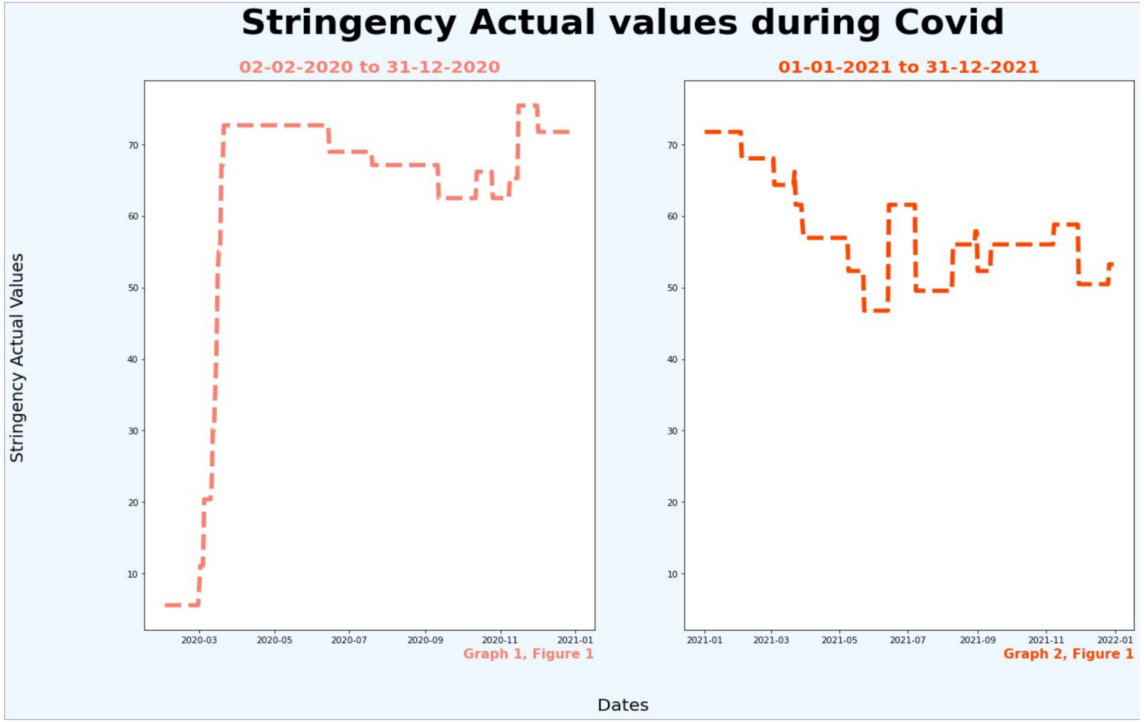
- **Covid and Covid vs Transactions:**
  - **Data Preprocessing**
  - **Visualisations & analysis**
- **Crime vs USA Population and Crime vs Transactions:**
  - **Data Preprocessing**
  - **Visualisations & analysis**

Data pre-processing:
COVID & Transactions

1. Importing data
2. Merging data sets
3. Rebuilding missing data
4. Standardisation
5. Normalisation
6. Deduplication
7. Verification and enrichment
8. Exporting data

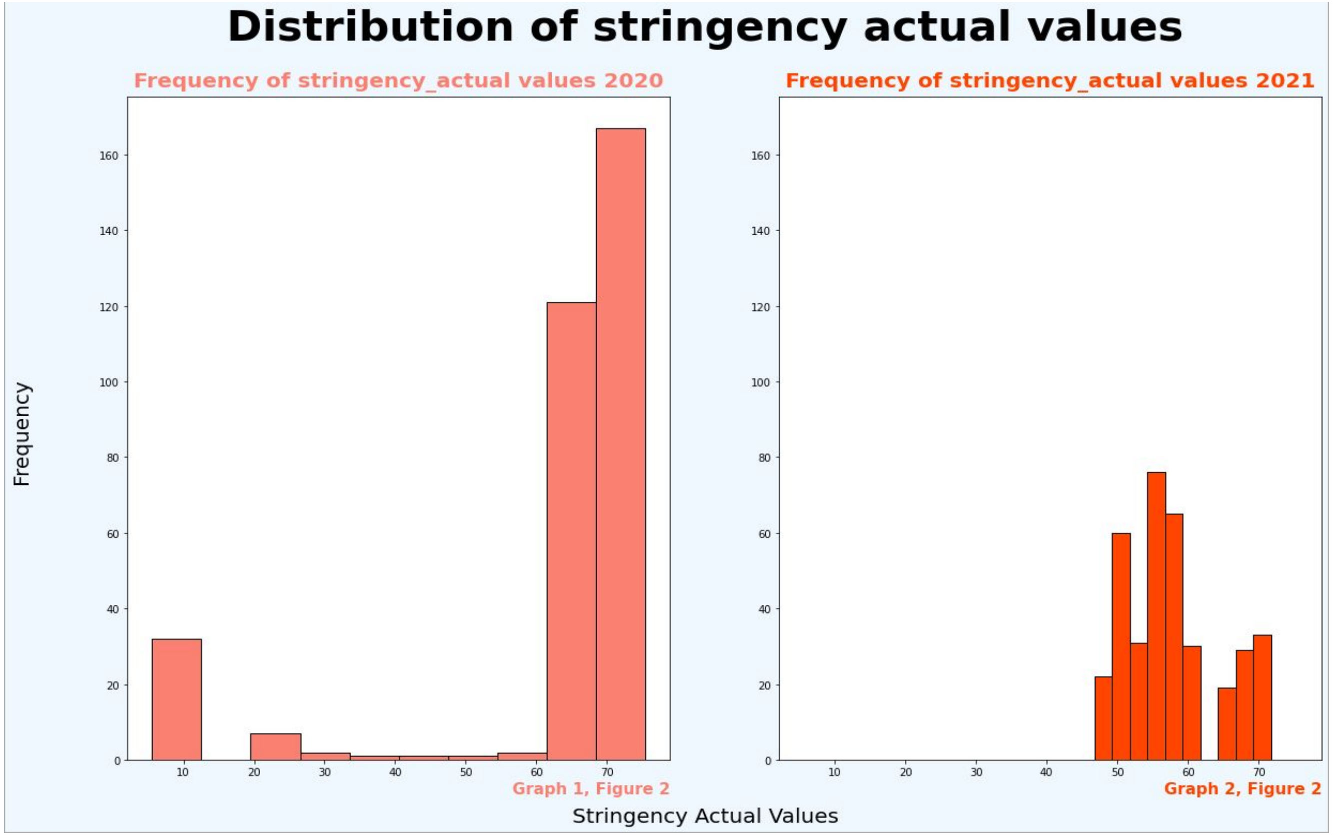


COVID19 Data Visualisations

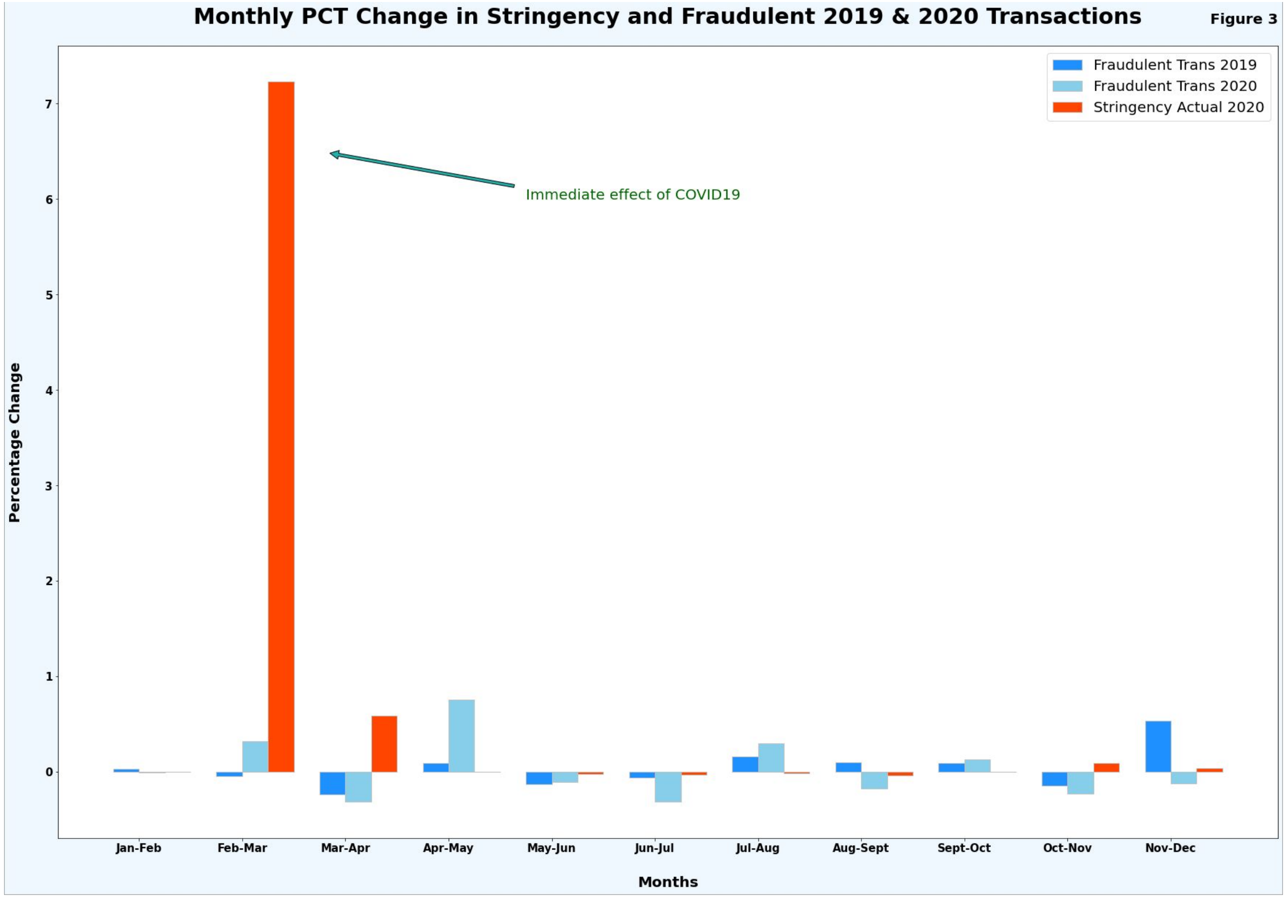


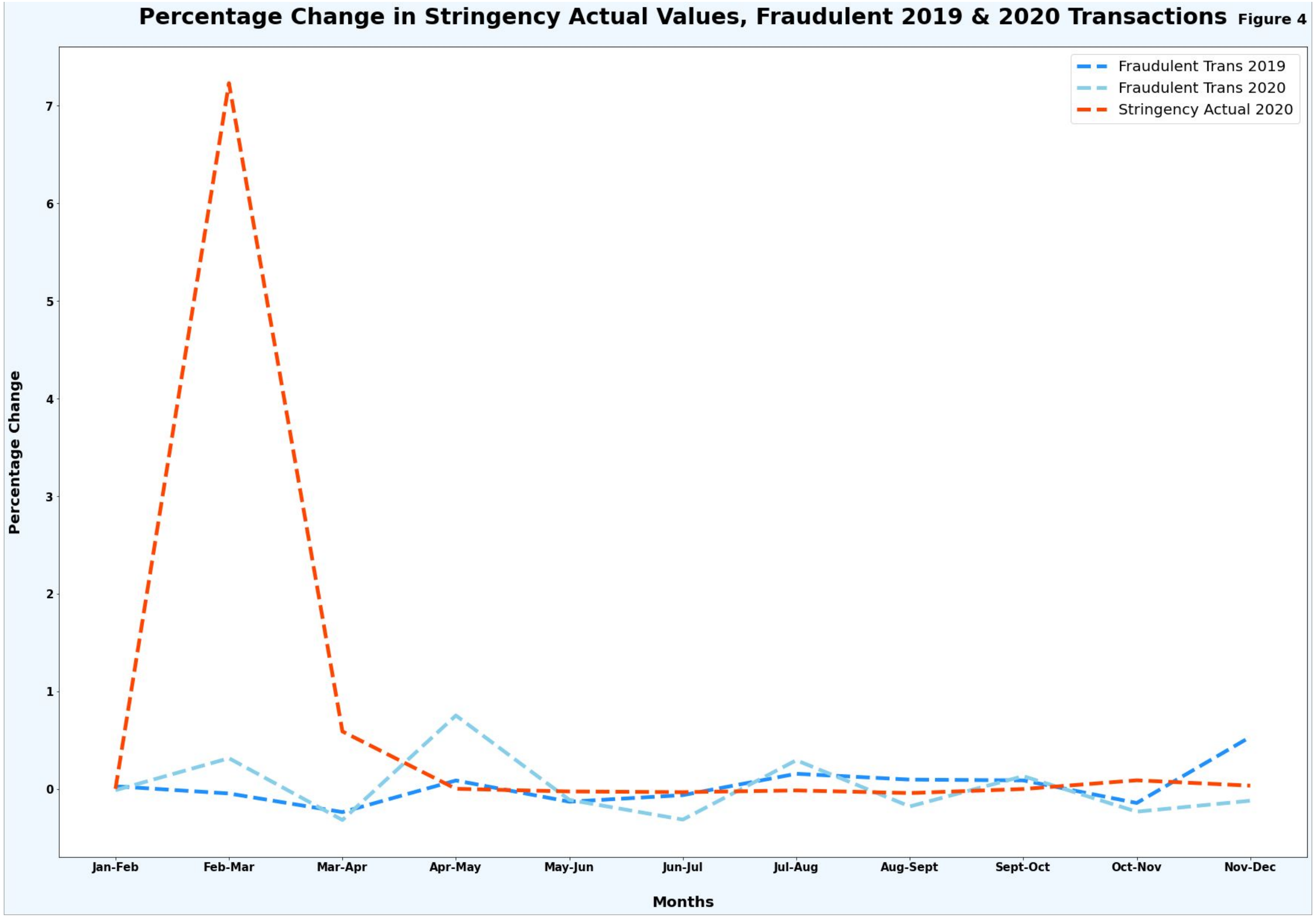
Changes in USA Stringency levels - coming to understand COVID19.

The difference in 2020 and 2021 stringency levels for the USA.



Covid vs Fraudulent CC Transactions



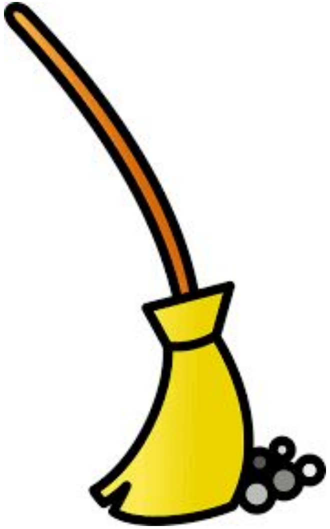


Percentage Change in Stringency Actual Values, Fraudulent 2019 & 2020 Transactions

Figure 4

Data pre-processing: Crime vs USA Population and Crime vs Transactions

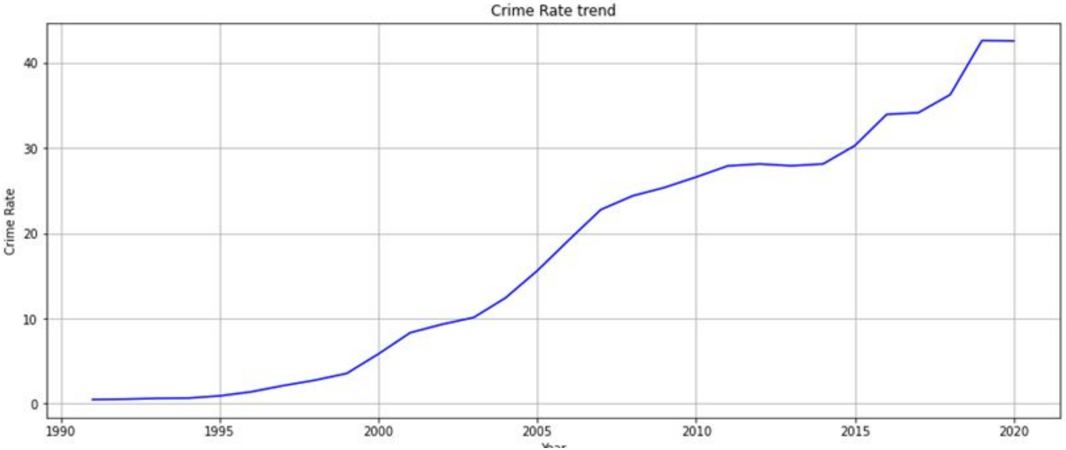
- Importing datasets and adjusting data types
- Dropping irrelevant columns and renaming other columns
- Merging datasets
- Building missing data (Year extraction from date)
- Calculating Crime Rates and Percentage Change
- Building graphs



CC and ATM fraud rate and US population

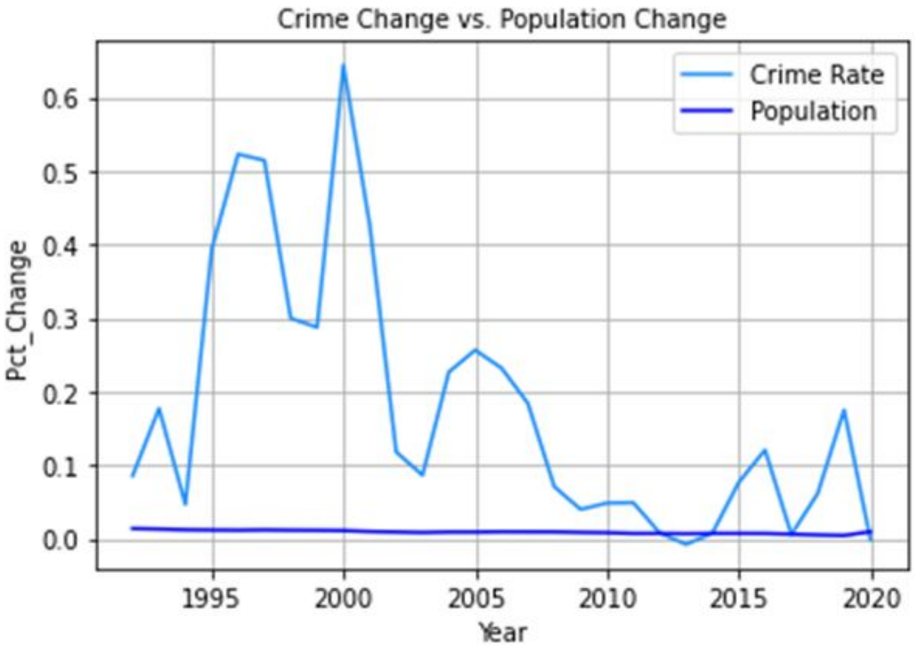
- Increased Crime Rate since 1991

Was the increase due to an increase in population?



CC and ATM fraud rate and US population

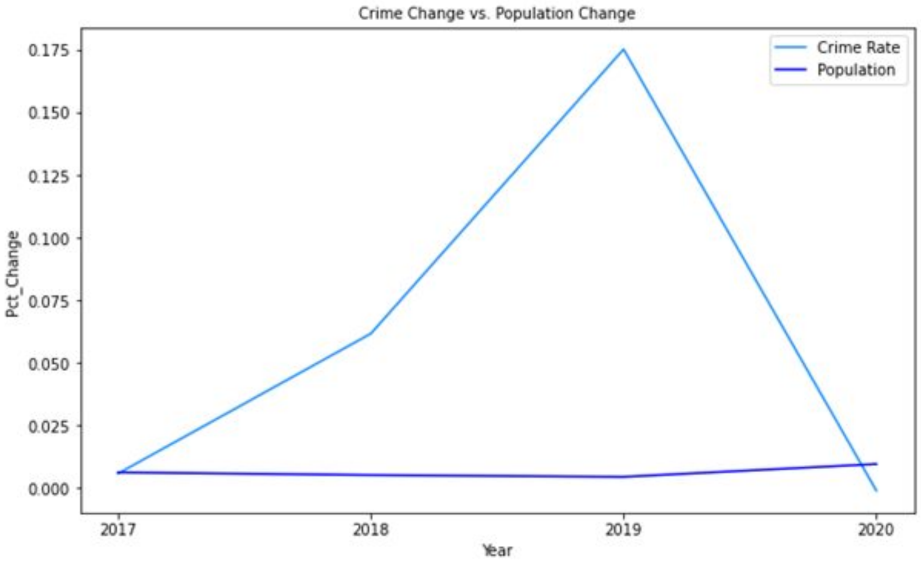
- Almost constant population
- Continuous fluctuation in crime rate



CC and ATM fraud rate and US population

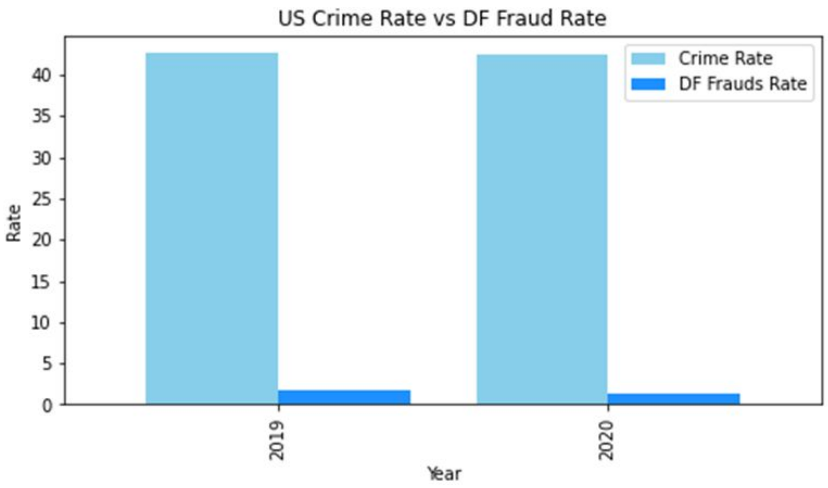
- Increase until end 2019
- Significant drop in 2020

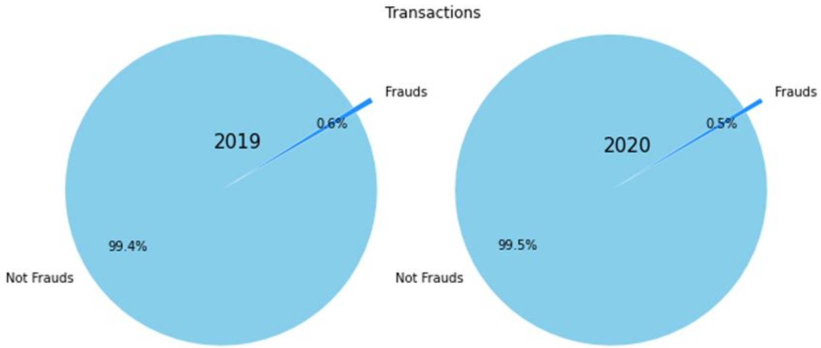
This could be an effect of 2020 lockdowns, but other data would be needed to proceed further...



Simulated Transactions

- ~0.5% frauds in the whole dataset





- >40 CC/ATM frauds vs. ~1 CC fraud
- great difference between the two rates

Conclusion: Were We Successful?

The focus question of this entire project was whether we were able to create a model that would predict if credit card transactions were fraudulent or not, this was something we were successfully able to achieve.

Both Juliana and Ilaria took different approaches in regards to model building, Juliana used the decision tree approach, and was successful.

Juliana was able to use the training data to learn patterns in the dataset and then apply that to new data. To ensure the model works, Juliana verified the models performance using precision and recall, and so a model was built in which credit card transactions could be detected for fraud

Conclusion: Were We Successful?

The sub-questions in relation to the focus question were answered by the data gathering and visualisation conducted by both Joey and Nicol. From the Crime API and Population Database:

- We were able to determine that fraud only counts as either 0.5% or 0.6% of all crime rates in 2019 and 2020, suggesting CC and ATM fraud is not as prevalent as we initially believed, this conclusion was drawn from the use of simulated data so cannot be said for certainty

- Were able to identify that there was no correlation between population change and crime rate, crime still grew whilst population change remained at a constant over a 30 year period.

Our final question was whether the effects of COVID19 restrictions result in more Credit Card Fraud than different periods of time throughout the year, after COVID19 API Analysis, we are able to see that there is no correlation at all between fraudulent transactions and the impact of stringency on the USA.

Conclusion: Were We Successful?

Of all the questions we answered, the one we were unsuccessful with was determining whether there were areas of higher fraud rates than other areas. The reason as to why we were unable to answer this question is because we were limited with our data.

The API used for CC and ATM offences only provided data at a Country level (US) whereas wanted to look into cities.

Despite this, we believe our project was successful as we were not only able to answer our focus question by creating a CC fraud detection model but also able to provide answers for our sub-questions in relation to fraud.