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

INTRODUCTION:

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 .

BACKGROUND: Questions and Analysis

WHAT QUESTIONS ARE WE TRYING TO ANSWER:

There are multiple questions we hope to answer through the project, the title question being the focus point with the sub-questions being answered through our findings. Questions are as follows:

- 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?

HOW CAN OUR ANALYSIS HELP:

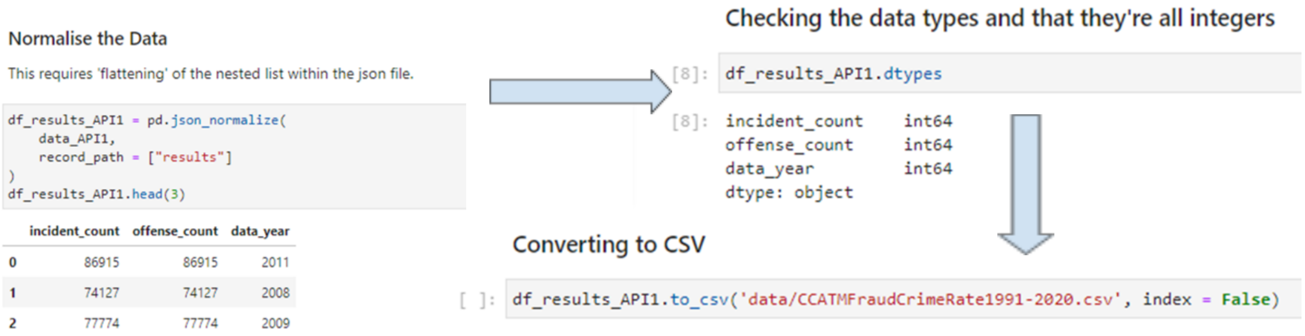
Although Credit Card Fraud analysis would be helpful to several industries, the target audience of the project would be Banks and Insurance Companies. We believe they would benefit the most from being able to detect fraudulent activities. Banks could possibly use the findings to prevent the transactions from occurring and insurance companies to better protect their customers from fraud and perform better risk analysis.

STEPS SPECIFICATIONS:-

Data Source 1: Credit Card Transactions Fraud Detection Dataset - Simulator by Brandon Harris
https://www.kaggle.com/datasets/kartik2112/fraud-detection?resource=download&select=fraudTrain.csv
Data Source 2: Oxford Covid-19 Government Response Tracker (OxCGRT) - (next referred as "COVID API")
https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker
Data Source 3: The FBI Crime Data API - (next referred as "CRIME API")
https://crime-data-explorer.fr.cloud.gov/pages/docApi
Data Source 4: US population - (next referred as "Population dataset")
https://fred.stlouisfed.org/series/POPTOTUSA647NWDB

Data Gathering - API: Nicol and Joey

Using the Crime API as an example; the API returns JSON or CSV data, organised around the data reporting systems the FBI UCR program uses. The API offers many endpoints from which to retrieve data from, among the many available endpoints, we decided to use the "**Victim Data Controller**" which provides victim demographic information for offences. The variable provided by this API were age, **count,** ethnicity, race, sex, relationship for each offense. We chose this endpoint as it provided data about offences divided by type of offence, one of which was "credit-card-automated-teller-machine-fraud", exactly what we needed. The data was at national level and we were unable to go in depth with analysis of the data based on the territory using the agencies, as only 63% agencies are participating in feeding the system with data. The response we got was a nested dictionary, and our data was the value of the key 'results'. To "flatten" our data we used '**pd.json_normalize()**' and then saved the entire dataset as a csv file.



Normalise the Data

This requires 'flattening' of the nested list within the json file.

```
df_results_API1 = pd.json_normalize(
    data_API1,
    record_path = ["results"]
)
df_results_API1.head(3)
```

	incident_count	offense_count	data_year
0	86915	86915	2011
1	74127	74127	2008
2	77774	77774	2009

Checking the data types and that they're all integers

```
[8]: df_results_API1.dtypes
```

```
[8]: incident_count    int64
     offense_count     int64
     data_year         int64
     dtype: object
```

Converting to CSV

```
[ ]: df_results_API1.to_csv('data/CCATMFraudCrimeRate1991-2020.csv', index = False)
```

Data Preprocessing - Ilaria & Juliana

Before starting the process of modelling, it is first necessary to understand what are the attributes involved in the dataset, what are the limitations of such attributes and what kind of statistical relationships we can find in the data.

Juliana - First, it was necessary to understand what were the columns involved in the dataset and what kind of data they contained. We found here that one of the columns contained date types, but the rest contained numerical variables or string. In order to proceed, it was also necessary to verify if there are null values in the dataset. In this process, we observed that there were no null values in the dataset. There were several adjustments in our dataset necessary for the modelling process. For instance, there is some information that could be able to identify a particular individual. In addition to that, there is information in our dataset that is too specific and could lead to model overfitting. For this reason, we have chosen to drop certain columns that could either cause ethical issues or damage the generalisation capacity of our model.

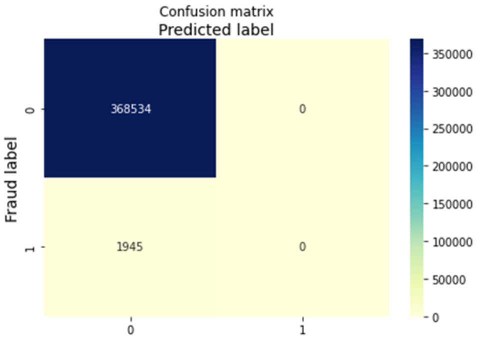
Encoding - Ilaria & Juliana

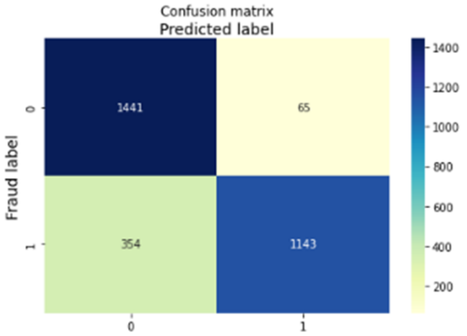
Juliana - One of the characteristics of our dataset is the amount of categorical variables. In this sense, we have decided to encode the categorical variables. The two techniques used for encoding were OneHotEncoder and Target Encoding. Juliana applied Target Encoding and Ilaria applied OneHotEncoding.

Ilaria - I decided to employ OneHotEncoder to encode data stored in the columns ‘category’, ‘state’, ‘gender’. This encoding method was selected as our categorical data does not have a pre-established order or continuity.

Predictive Modelling- Ilaria & Juliana

In order to train a machine learning model, it is necessary to divide the data into training data and testing data. This means that one part of the dataset will be used so that the algorithm can learn the patterns in the dataset. The testing part of the dataset is to verify if the model is able to correctly predict the labels. Our aim is producing a model that is perhaps less accurate overall, however more often correctly identifies fraudulent transactions (does not produce a lot of false positives or false negatives).





Ilaria - I created a OneHotEncoder pipeline and employed it to train a logistic regression model, while Juliana used Target Encoding to train a decision tree model. As shown in the Confusion Matrix on the right, the logistic regression model is currently prone to overfitting. Although its accuracy is extremely high, it can only correctly label genuine transactions, while it has labelled all fraudulent transactions incorrectly. The overfitting issue is caused by the unbalanced nature of the dataset. Currently, the model has an abundance of examples for genuine transactions, however it has very few opportunities to learn about the characteristics of fraudulent transactions.

I tried experimenting with resampling techniques to deal with the unbalanced nature of the dataset and overcome the logistic regression model overfitting issue. I first used upsampling to artificially inflate the instances of fraud to match those of genuine transactions. This however created duplicate data of fraudulent transactions, which resulted in another logistic regression model prone to overfitting. Then I applied downsampling to only select a smaller pool of genuine transactions to match the number of fraud instances. This initially showed very encouraging metrics, as demonstrated in the confusion matrix on the right. However, after restarting the runtime multiple times to test whether different cuts of the downsampled majority class would produce consistent results, it became clear that performance metrics are very inconsistent when using downsampling techniques on this dataset.

Finally, I decided to test cross-validation and cost-sensitive learning techniques to make another attempt at harvesting logistic regression models’ potential for fraud prediction. Both these methods produced metrics demonstrating that logistic regression continues to be prone to overfitting, with high accuracy scores, but low recall and precision scores. In conclusion, employing cross validation techniques has not revealed any promising path forward to successfully prevent logistic regression models from overfitting. It appears that employing a decision tree model to predict fraudulent transactions is the best course of action.

Juliana - For the Decision Tree model with Target Encoding, the model offers better precision and recall metrics, as we will show in the result reporting section.

IMPLEMENTATION AND EXECUTION:

Development of the project was done through task delegation, tasks were assigned to each member of the project, where members chose tasks they were most interested in. Task delegation was as so:

Modelling Taskforce (2 people) - Juliana & Ilaria

Research best classification models for fraud prevention, preprocess and encode data, split dataset into train and test, fit the model into the data, analyse results (at least two different models).

API & Data Visualisation Taskforce (2 people) - Nicol & Joey

Find an open access API that could complement our analysis on credit card fraud (example: crime, weather). Request data from the API and communicate with the visualisation taskforce to integrate it into visualisation. Perform meaningful exploratory analysis with data, using data to tell a story and convey an idea. (Example: plot map of credit card fraud, show main institutions involved, etc). Initially API and Visualisation were separate tasks however, the group decided it would be better for the project if Nicol and Joey were to work together, as a pair.

Project Management Taskforce (1 person) - Rumaanah

Using an Agile approach, make sure everyone is doing their tasks, help colleagues if anyone has any issues, technical or not, keep track of the work, create the report and presentation, and also worked on data analysis of data visualisations created for the report. Made sure the parts of the project fit together. Point of contact in the communications with Polly.
Despite the individual task assignment, collaborative effort was at the core of the project as it was mutually understood to be the most important factor of a successful project.

Tools and Libraries Used:

Pandas, Numpy, Matplotlib, Python, APIs, SciKit, Machine Learning, Jupyter Notebook, Seaborn.

Implementation Process: Achievements

Nicol: As the project I gained confidence in coding, using the tools and also approaching data analysis so deeply involved in the process. Also the team work went well: a team of motivated and resourceful people who cooperated and helped each other.

Joey: Working on this project has given me the confidence in coding to retrieve APIs and create visualisations to a standard that I am more comfortable with. I better understand the steps and processes required and how to better clean and prepare my data.

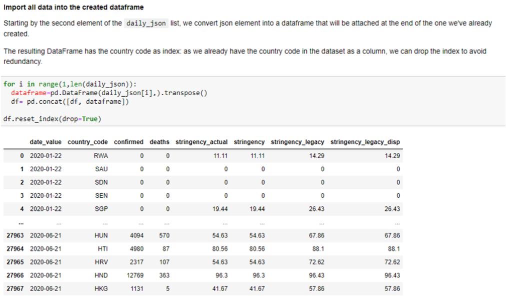
Rumaanah: This was my first time as acting Project Manager and was initially worried that my role wouldn't be very hands on, however I completely underestimated how much organisation,planning, and running around is required, it is role I am glad I took on and one of my proudest achievements was narrowing down this report from 30+ pages to 6 pages whilst still ensuring all the requirements were met.

Implementation Process: Challenges

Rumaanah: One of the first problems we had to deal with was when we wanted to use a real banking data set for the credit card transactions, although we found a dataset, it was unusable as vital data had been blanked out for protection of the customers; we were able to find a way around it by using a simulated data set.

Joey: A lot more research is required for obtaining the correct API than I initially expected. This is to ensure it provides you with the information you seek from a reliable source. The selection of visualisations was also vital in how we wanted to depict the analysis we gained from the datasets.

Nicol: The COVID API result was a nested dictionary and was complicated to get to the data inside, to extract the data properly, we decided to split the json by date, create a dataframe with the data of the first day, with an iteration we extracted the other data one by one, converted into dataframe and concatenate them to the first dataframe, as seen on the image to the right. The crime API also had nested dictionaries, and didn't have any documentation to explain the data in detail, so we couldn't use it to calculate the overall crime rate of the US.



Import all data into the created dataframe

Starting by the second element of the `daily_json` list, we convert json element into a dataframe that will be attached at the end of the one we've already created.

The resulting DataFrame has the country code as index: as we already have the country code in the dataset as a column, we can drop the index to avoid redundancy.

```
for i in range(1,len(daily_json)):
  dataframe=pd.DataFrame(daily_json[i],).transpose()
  df= pd.concat([df, dataframe])

df.reset_index(drop=True)
```

	date_value	country_code	confirmed	deaths	stringency_actual	stringency	stringency_legacy	stringency_legacy_disp
0	2020-01-22	RWA	0	0	11.11	11.11	14.29	14.29
1	2020-01-22	SAU	0	0	0	0	0	0
2	2020-01-22	SDN	0	0	0	0	0	0
3	2020-01-22	SEN	0	0	0	0	0	0
4	2020-01-22	SGP	0	0	19.44	19.44	26.43	26.43
...	...	...	...	...	...	...	...	...
27963	2020-06-21	HUN	4094	570	54.63	54.63	67.86	67.86
27964	2020-06-21	HTI	4980	87	80.56	80.56	88.1	88.1
27965	2020-06-21	HRV	2317	107	54.63	54.63	72.62	72.62
27966	2020-06-21	HND	12769	363	96.3	96.3	96.43	96.43
27967	2020-06-21	HKG	1131	5	41.67	41.67	57.86	57.86

Implementation Process: Decision to Change Something

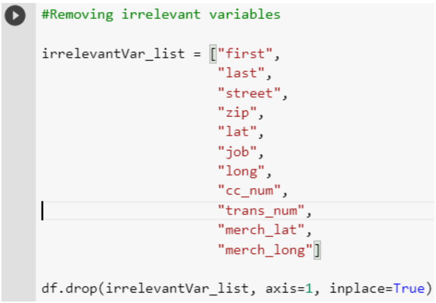
Rumaanah: There were hopes that a Weather API (Weather API Documentation https://rapidapi.com/visual-crossing-corporation-visual-crossing-corporation-default/api/visual-crossing-weather) may also be used as a factor to determine the effect of weather on credit card fraud however the data proved difficult to retrieve, as there was no way around it without paying for the data.

Joey: As noted in the challenges, the graph type choices were difficult to make at the beginning without cleansing and preparing the data first. The decision to make bar plots and line graphs as opposed to a scatter graph for the comparison of stringency v fraudulent transactions became more apparent as the data was leading us in that direction. To avoid creating bias from visualisation depictions we changed the graph type.

Ilaria - Rather than change something, I would like to further experiment with cross-validation techniques and cost-effective learning. I do not think I have exhausted the potential of these techniques, however given the limited resources and time I could not perform any further tests

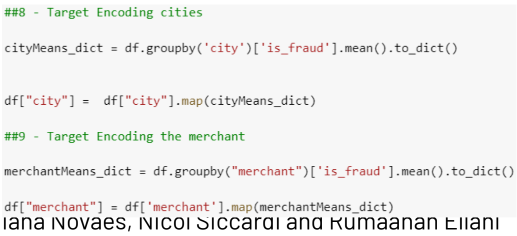
RESULT REPORTING - Predictive Modelling:

Juliana - Predicting fraud is a binary classification problem, meaning that there are two possible outcomes: the transaction is a fraud (1) and the transaction is not a fraud (0). With Ilaria trying logistic regression, I decided to try a decision tree as an option, since they are models based on different mathematical paradigms. There were several adjustments in our dataset necessary for the modelling process, including encoding and cleansing.



```
#Removing irrelevant variables

irrelevantVar_list = ["first",
                      "last",
                      "street",
                      "zip",
                      "lat",
                      "job",
                      "long",
                      "cc_num",
                      "trans_num",
                      "merch_lat",
                      "merch_long"]

df.drop(irrelevantVar_list, axis=1, inplace=True)
```

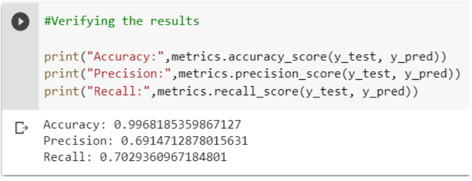
In order to train a machine learning model, it was necessary to divide the data into training data and testing data. The testing part of the dataset was to verify if the model is able to correctly predict the labels. As mentioned I applied decision tree, which ended up presenting good metrics. Decision tree which is a model that includes conditional 'control' statements to classify data, starting at a single point (or 'node') which then branches (or 'splits') in two or more



```
##8 - Target Encoding cities

cityMeans_dict = df.groupby('city')['is_fraud'].mean().to_dict()


df["city"] =  df["city"].map(cityMeans_dict)

##9 - Target Encoding the merchant

merchantMeans_dict = df.groupby("merchant")['is_fraud'].mean().to_dict()

df["merchant"] = df['merchant'].map(merchantMeans_dict)
```

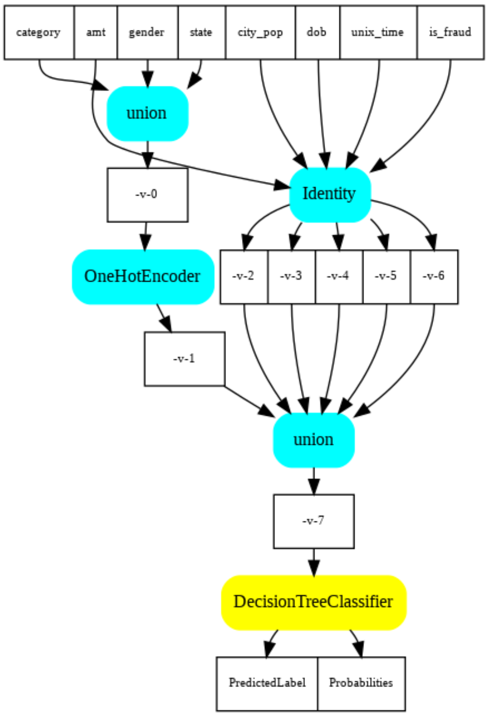
directions. Each branch offers different possible outcomes, incorporating a variety of decisions and chance events until a final outcome is achieved.

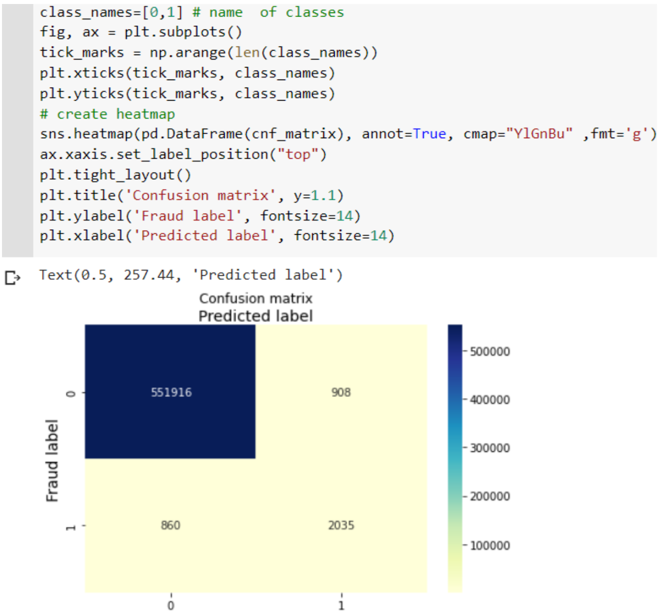
After applying the model to the data, an important step is to verify its performance, to evaluate its capacity for generalisation. Considering the nature of our dataset, which is extremely unbalanced, precision and recall was used as a metric for evaluating the model, this is because it is able to evaluate not only the quantity of correct predictions vs incorrect ones, but also look at the number of false positives and false negatives.



```
#Verifying the 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: 0.9968185359867127
Precision: 0.6914712878015631
Recall: 0.7029360967184801
```





```
class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Fraud label', fontsize=14)
plt.xlabel('Predicted label', fontsize=14)
```

```
Text(0.5, 257.44, 'Predicted label')
```

RESULT REPORTING - Data Visualisation of APIs:

Nicol - Crime API and Population Database Analysis: Before getting into the comparison between the number of the frauds and the crime rate, we considered it useful to show the trend of the Crime Rate Limited to CC (Credit Card) and ATM Fraud in the time period 1991-2020, as depicted in the graph below. The number of offences related to 2020 must be considered as the total number of offences reported by the end of 2020. As our crime dataset is the result of the offences reported to the FBI by the agencies, with only 63% Agencies sending data, we must consider our dataset as a sample of the real number of offences that occurred. We, also, cannot be sure that the offences occurring in the competence area of these agencies correspond to the overall number of offences. Hence, for our considerations, we assumed that their data reflects the behaviour of the overall population. We plotted our crime data to see the trend of this index.

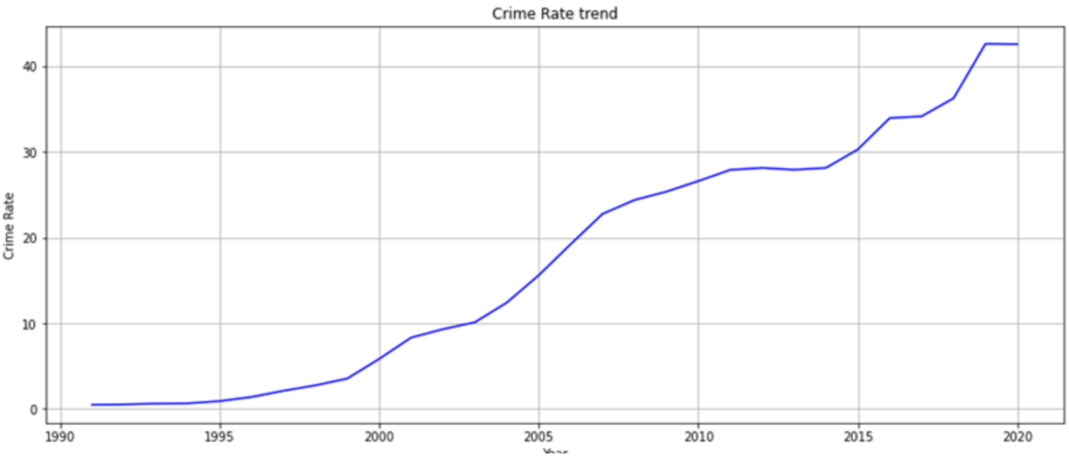


Figure 1: From this graph we can see a steady incline in crime rates over the years, between 1991 to 2010, crime rose to 28% and in 2010 to 2020, crime increased from 28% to 43%, averaging crime rate at an approximate growth of 14% per decade.

We should next question the trend in population, was the increase in crime rate due to an increase in population? To understand whether the rate of population was a factor affecting the crime rate, a data visualisation depicting percentage change of both was created, the results of which can be found below.

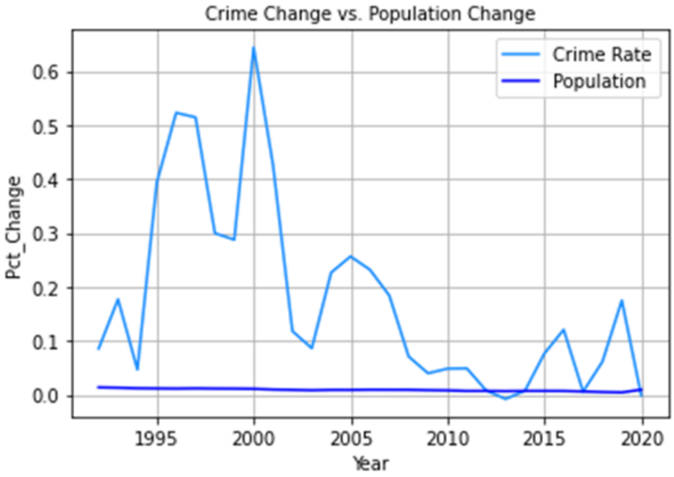


Figure 2: From this line graph we see both crime rate and population, the population line remains at a constant with hardly any change over the almost three decade period, unlike the crime rate measurement. Although there was a continuous fluctuation in crime rate, with the highest point at 2000 and the lowest at 2012, the rate of population did not change. It is therefore safe to say the rate of population had no effect on the rate of crime.

One of our questions was related to the hypothetical increase of fraudulent CC transactions as a result of lockdowns due to COVID19. As we can see in the graph below, in the US, Crime rate related to CC and ATM frauds did in fact increase in 2018 and 2019 but did decrease in 2020: this could suggest that there may have been a potential affect by COVID19, however lockdowns did not occur until 2020 onwards and the data we retrieved did not pass 2020.

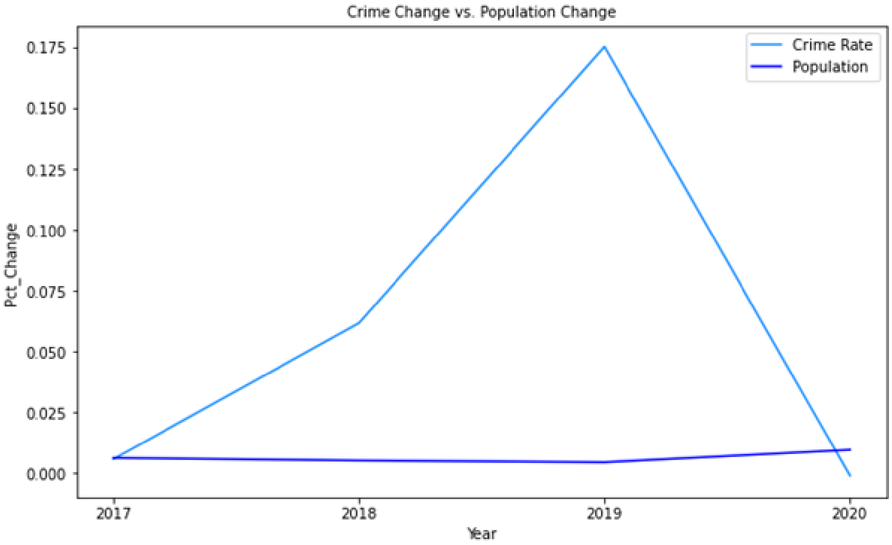


Figure 3: This line graph focuses on a 3 year period between 2017 to 2020, again it depicts the percentage change of Crime Rate and Population Change. As previously mentioned, population change has remained at a constant level at approximately 0.015%. It is evident to see the consistent change in crime rate however what is most interesting is the highest peak being at the end of 2019 which is then followed by a significant drop. As mentioned earlier this may suggest COVID19 lockdowns might have had an effect on crime rates however data does not surpass 2020 so it can not be said for sure.

We then calculated the crime rate types, 'Not Fraud' vs 'Fraud' as a percentage for both 2019 and 2020. We decided to use a pie chart to show the data composition because a bar chart would have had a very high bar for legitimate transactions and almost just a line for fraudulent transactions, which provided the following result:

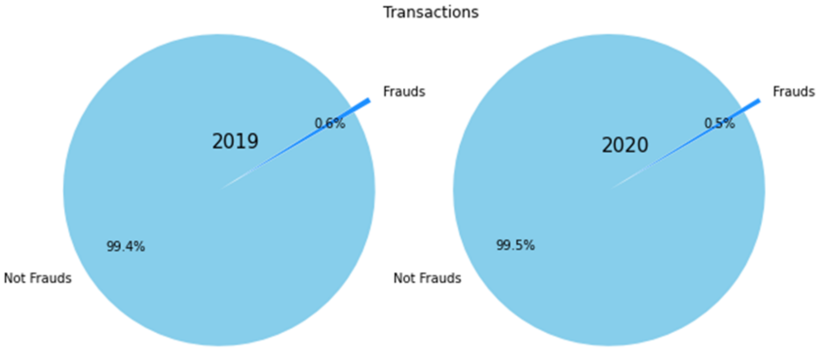


Figure 4: As we can see, 99.4% and 99.5% of transactions respectively in 2019 and 2020 were legitimate, and we only have 0.6% of transactions in 2019 and 0.5% of transactions in 2020 classified as fraud, this could further prove our hypothesis that COVID19 did have an affect on fraud crime as we saw a drop by 0.1% however there is not enough data to back this claim.

A direct comparison of Crime Rate to Fraud Rate was then conducted using the same assumption we used for our simulated data set, that the dataset in use is a representation of all CC users in the US , and calculate the CC Fraud Rate with the same formula, resulting in the number of fraudulent transactions every 100k citizens.

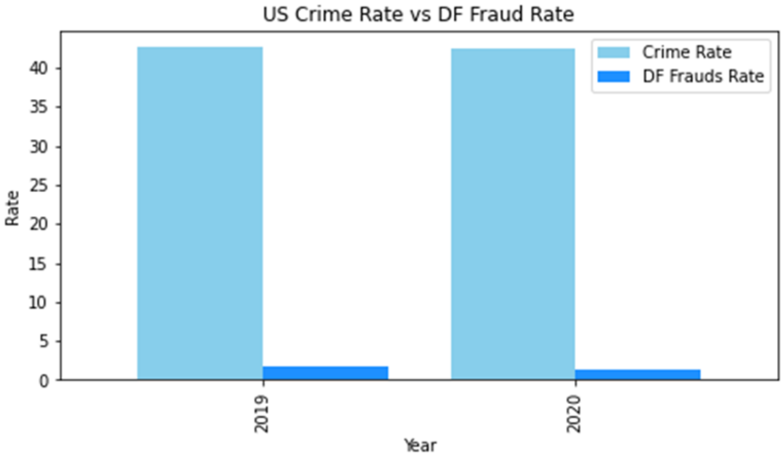


Figure 5: In both 2019 and 2020 we have a rate of more than 40 CC/ATM frauds every 100k citizens in the US, in our dataset the Frauds Rate is slightly above 1, but the offences count used to calculate the crime rate comprehends CC AND ATM frauds: we don't have enough data to understand in which proportion.
Looking at our graph we could say that the Fraud Rate reflects in some way the Us Crime rate for CC and ATM fraud. However, the great difference between the two rates in both years make it clear that we don't have enough data and are not good enough to proceed further with the analysis and formulate a significant hypothesis.

Joey - COVID19 API Analysis: Before we start the comparison between stringency percentage change for USA during COVID v. fraudulent transactions percentage change in 2019 & 2020. We found it important to have an understanding of how the stringency values did change over the course of 2020 and 2021 in the USA. These 2 figures endeavour to provide some insight from the data we have to give us an understanding of what the US government policies were like. We have used the stringency_actual values as opposed to the other calculated ones as we do not know all the initial values for which they were used for the calculations.

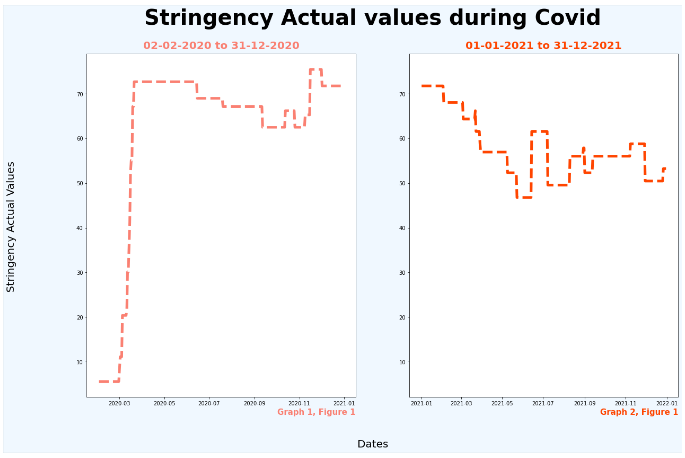


Figure 1: the sharp rise in stringency values in 2020-03 indicates when COVID19 first impacted the USA government's policies. From these we are able to see the dips and rises according to what you can infer as seasonal changes. There is a slight increase of stringency in 2020-11 at the beginning of winter, then as winter draws to an end the stringency values slowly decrease. This is not to say they are gone completely. With COVID19 there was a lot of uncertainty and even though Summer came around 2021-05/06 with stringency being at the lowest levels for the past year, it sharply increased back to 63 possibly due to the increase in number of COVID19 cases.

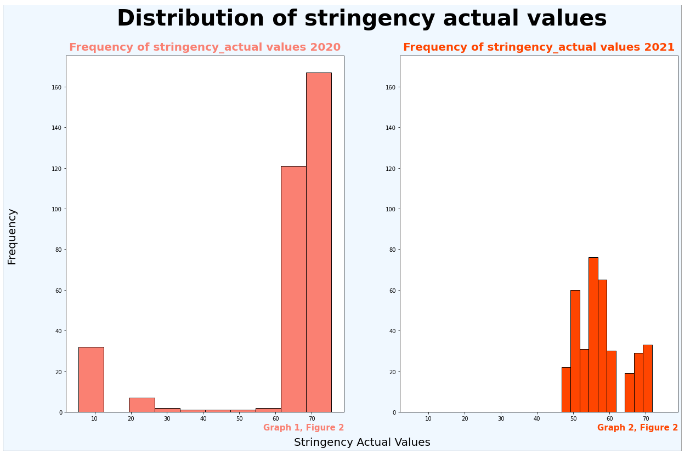


Figure 2: A comparison of the distribution of stringency actual values between 2020 and 2021. This is to see how the USA changed their stringency as we got to understand COVID19 better. They share the same x and y axes so you can see the difference in values immediately. With these graphs it is clear to see that consistently higher levels of stringency were present for the year of 2021. There is a spike for the year 2020 but there was more uncertainty and so that is why there are values towards the lower end of the spectrum.

We then get to answering the question, 'did the effects of COVID restrictions result in more Credit Card Fraud than different periods of time throughout the year?'.

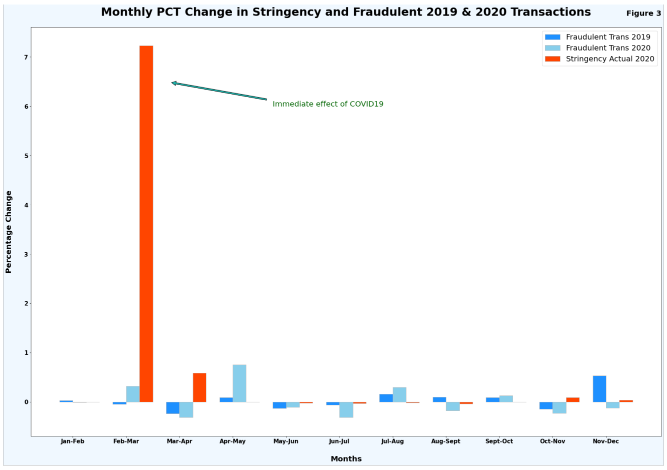


Figure 3: we are able to see a major spike in stringency percentage change for the months of February to March. We do not see this same effect for fraudulent transactions in 2020 as one would expect with everyone being indoors. This could be due to the dataset being simulated and not real-world data.

From what we can infer from the datasets, there is no correlation at all between fraudulent transactions and the impact of stringency on the USA.

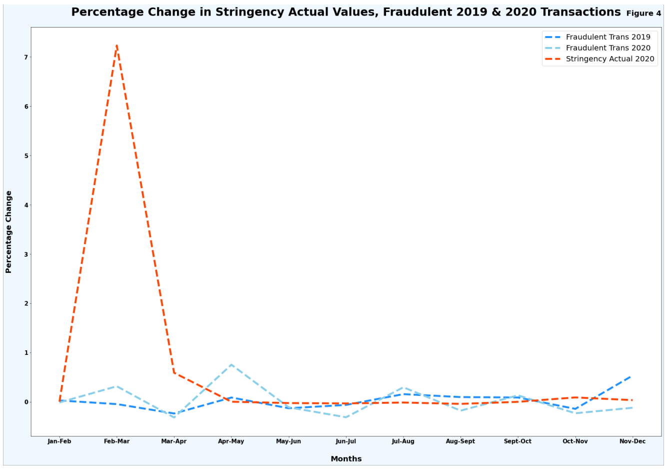


Figure 4: depicts the same lack of correlation between transactions and stringency. This visualisation was created to highlight this in a more effective way. It is from creating multiple graphs which we are able to determine which would best depict the information the data is trying to show us.

CONCLUSION

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, Ilaria took on linear regression and Juliana used the decision tree approach, after continuous attempts from both, Juliana was successful. Found under 'RESULT REPORTING - Predictive Modelling' 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.

The sub-questions in relation to the focus question were answered by the data gathering and visualisation conducted by both Joey and Nicol, the first being what are the crime rates in comparison to fraud rates. This question was of interest to truly understand how big fraud is as a type of crime, and whether or not population change affects this. Looking at figure 4 from Crime API and Population Database Analysis 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 in comparison to other crime types, however this is simply the percentage of fraudulent transactions in our simulated dataset and we can't draw that conclusion for certain. While looking at figure 2 we identified 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 COVID restrictions result in more Credit Card Fraud than different periods of time throughout the year, looking at just figure 3 and figure 4 of the 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.

Of all the questions we answered, the one we were unsuccessful with was determining where 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.