**Auto Insurance Claims- Fraud Detection**

**Problem Statement:**

Insurance fraud is a huge problem in the industry. The automobile insurance fraud is one of the main challenges for insurance companies as it is very difficult to identify fraud claims. This leads to greater financial losses, which results in hike of insurance premium prices annually. This kind of fraud ranges from misrepresenting facts on insurance papers and inflating insurance claims to staging accidents and submitting claim forms for injuries or damage that never occurred. Because of such activities loyal customers are also affected, as it results in additional time and review before insurers pay legitimate claims.

Machine Learning is in a unique position to help the Auto Insurance industry with this problem. The aim of this project is to build a ML model on the basis of dataset we have, that can detect if an insurance claim is fraudulent or not. But there is a challenge behind fraud detection in machine learning because frauds are far less common as compared to real insurance claims. This type of problem is known as imbalanced class classification. By building a model that can detect auto insurance fraud, will help the auto insurance companies to reduce losses. I’ll be testing several different classification models on the dataset.

**Data Analysis:**

DataSource: <https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Automobile_insurance_fraud.csv>

The dataset has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made. The dataset consists of 1000 samples with 40 different parameters including the target variable (fraud reported). The columns are as follows:

months\_as\_customer

age

policy\_number

policy\_bind\_date

policy\_state

policy\_csl

policy\_deductable

policy\_annual\_premium

umbrella\_limit

insured\_zip

insured\_sex

insured\_education\_level

insured\_occupation

insured\_hobbies

insured\_relationship

capital-gains

capital-loss

incident\_date

incident\_type

collision\_type

incident\_severity

authorities\_contacted

incident\_state

incident\_city

incident\_location

incident\_hour\_of\_the\_day

number\_of\_vehicles\_involved

property\_damage

bodily\_injuries

witnesses

police\_report\_available

total\_claim\_amount

injury\_claim

property\_claim

vehicle\_claim

auto\_make

auto\_model

auto\_year

fraud\_reported

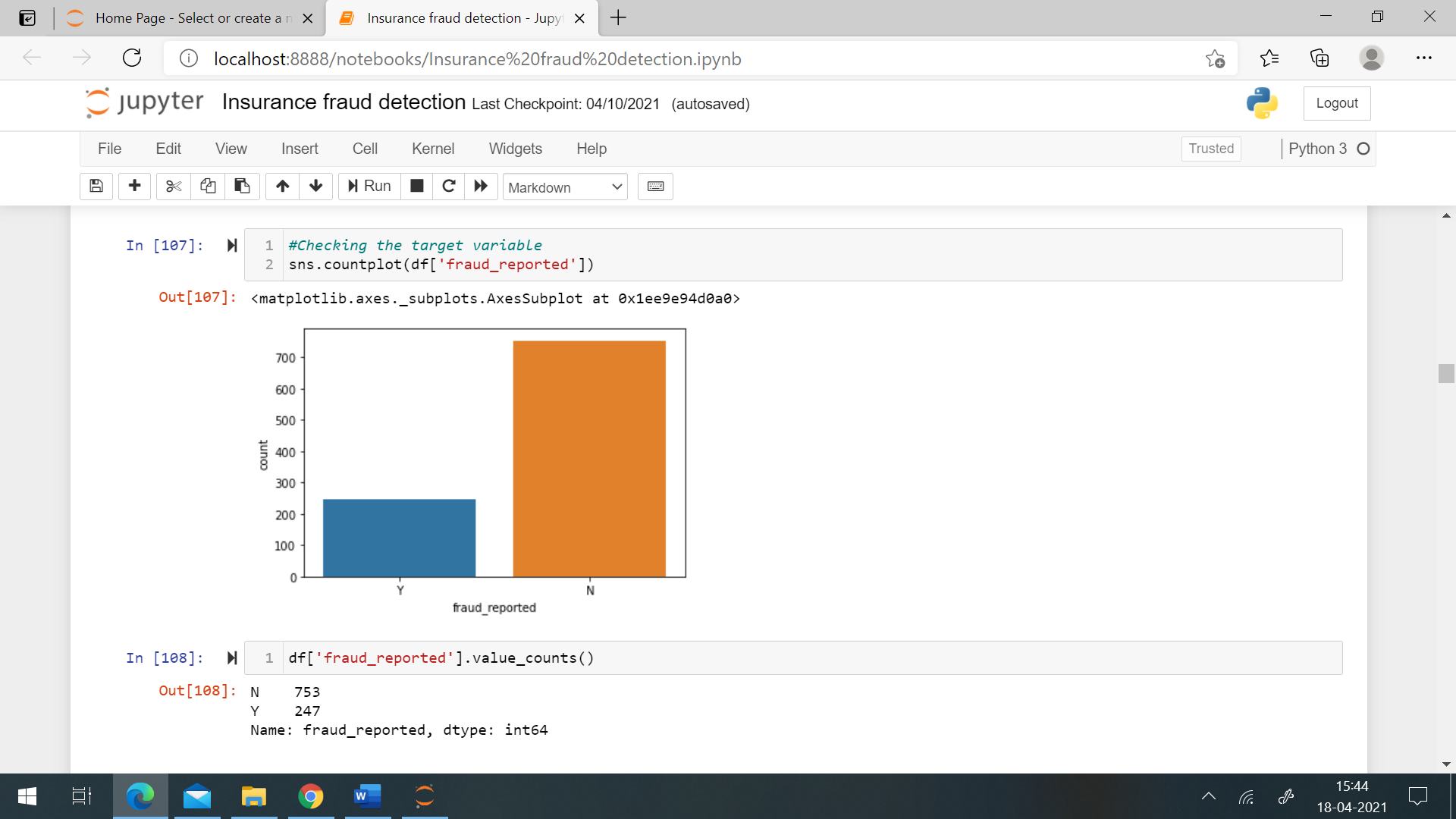
\_c39

The last column (\_c39) is null, it doesn’t contain any information. Therefore, we have 39 columns with information. The target variable (fraud reported) is categorical variable, it contains Y or N. Y is for fraud claim reported and N is for fraud not reported. It doesn’t contain the name of insurance company so, it is difficult to say if the data is from multiple companies or just one. The data set consist of incidents which occurred from January 2105 – march 2015. The data set contain policies of three states only. And there are no missing values in our data set other than \_c39 column.

**Exploratory Data Analysis:**

Target variable: -

In the data set, fraud\_reported is the target variable. There are 247 fraud cases and 753 non fraud cases (24.7% fraud cases and 75.3% are not fraud). This shows that our dataset is imbalanced.



Independent variables: -

Most of the customers are aged between 39-43 years and majority of customers are female. All the insurance claims are form Ohio, Illinois and Indiana, these are the name of states located in United States. Majority of claims are from Ohio (OH) followed by Illinois (IL). The deductible amount i.e., the amount paid out of pocket by the policy holder before an insurance provider pay any expenses, is either 500 or 1000 or 2000. Umbrella limit is zero for most of the customers, it is the limit which covers eligible damages that exceeds the auto insurance limits.

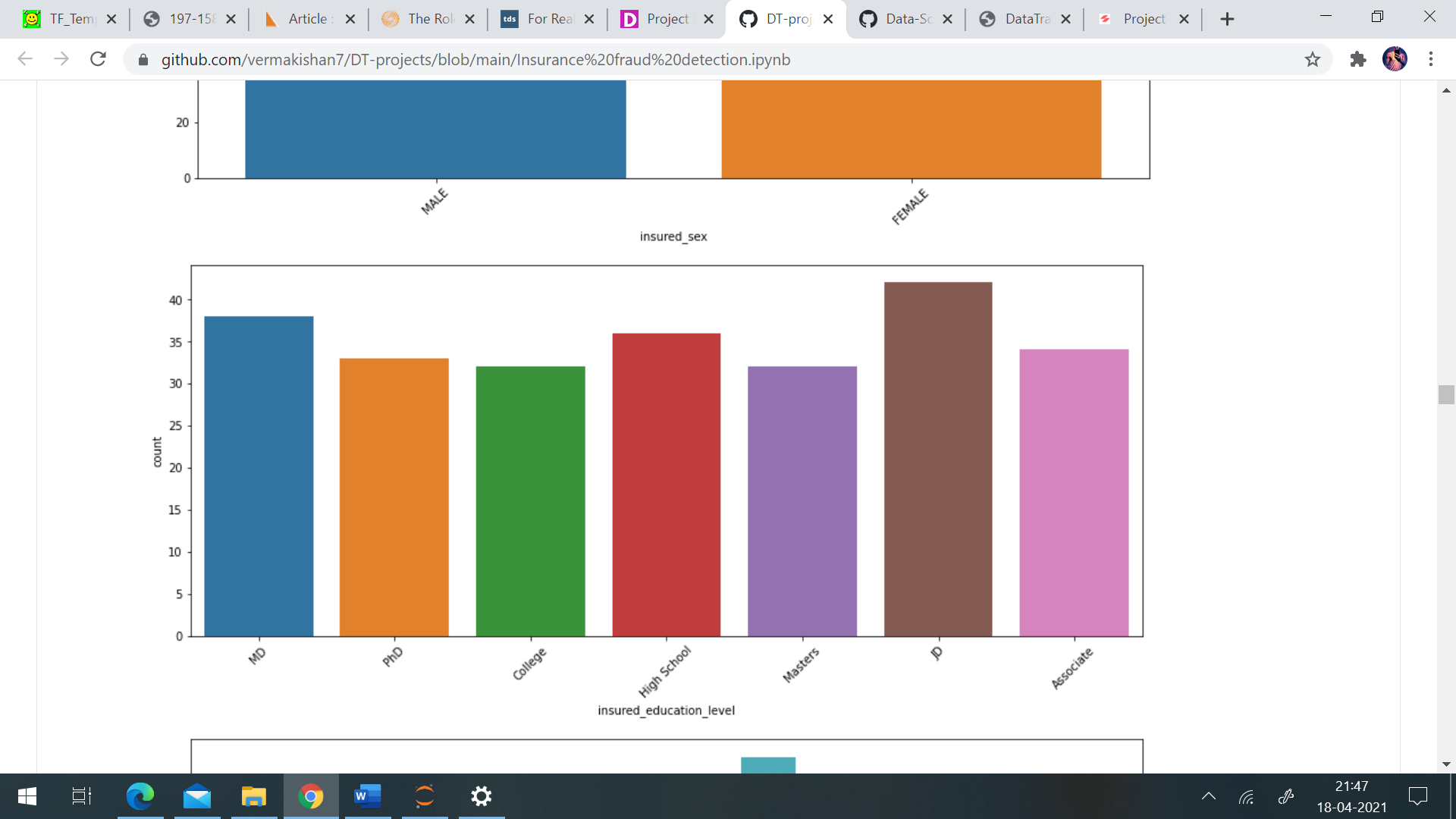
The education level of majority of customers in the dataset is high school or JD (Juris Doctorate), JD is the highest education available in the legal profession in the United States, it shows that all the customers are well educated. The occupation of majority of customers in the dataset is machine operation inspector followed by prof-speciality. Talking about the hobbies of the customers, reading and exercise tops the list and least are interested in cross-fit and basketball. Most of the customers are married and have their own child, there are less customers who are unmarried.

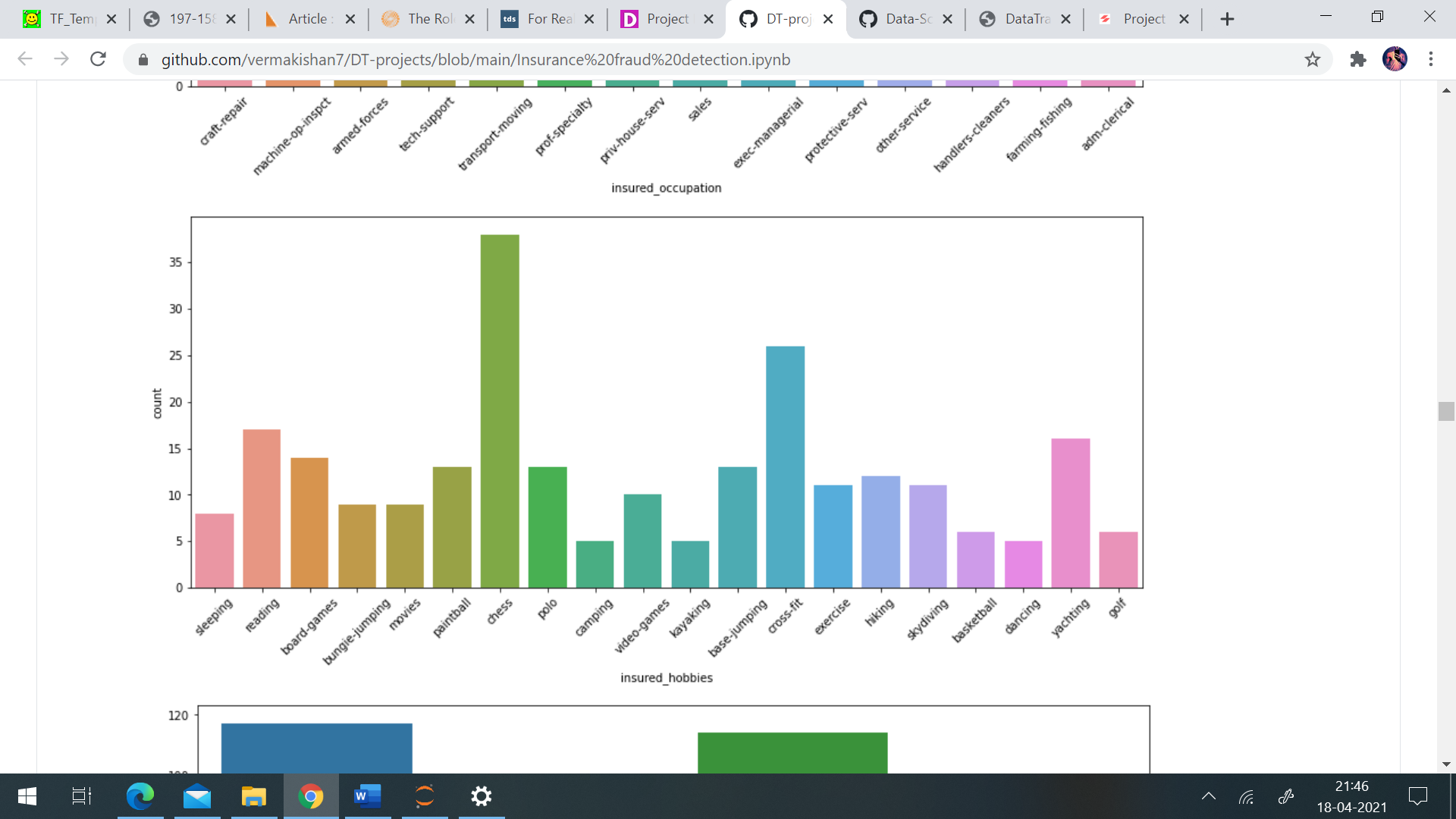


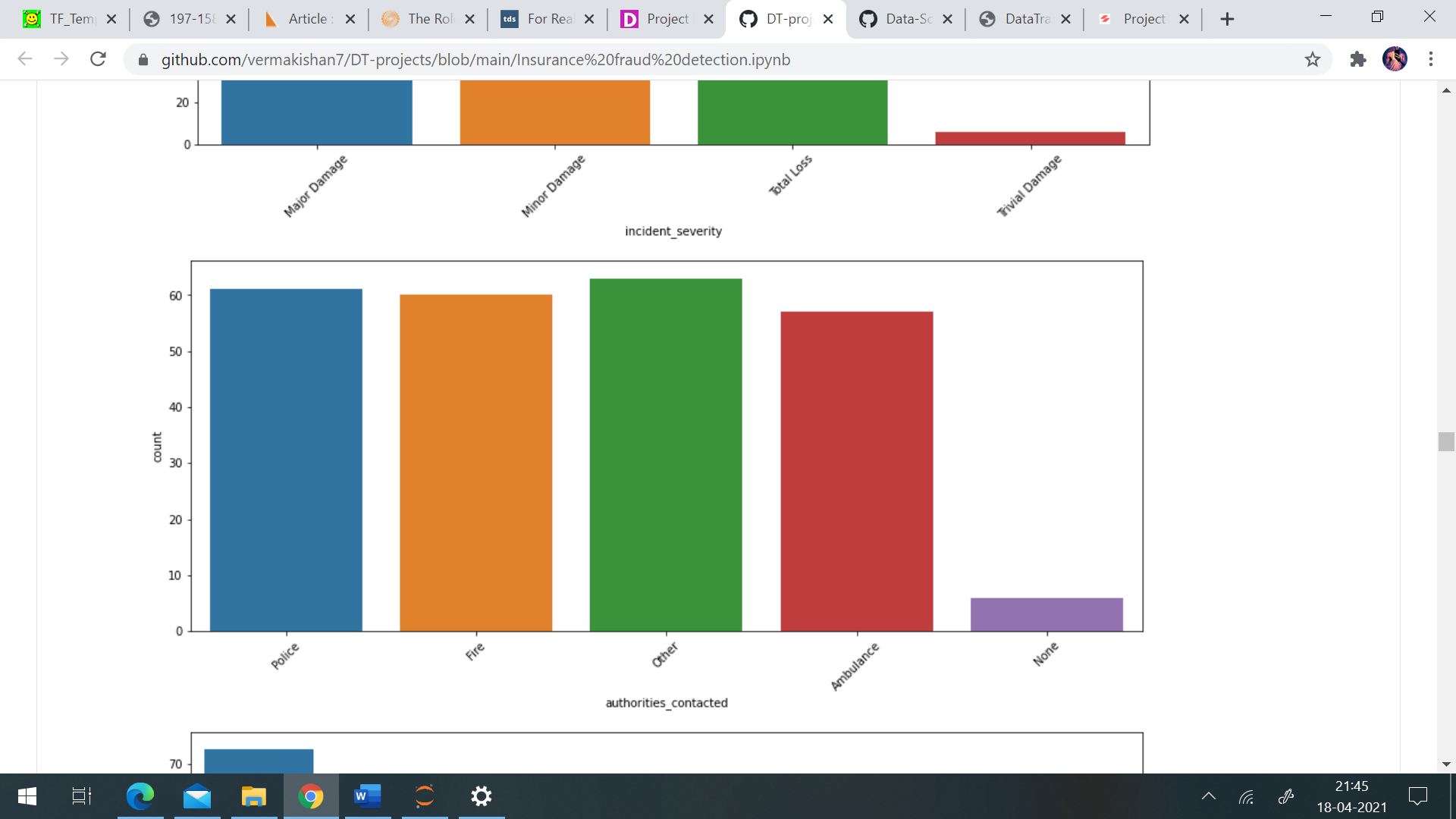
Coming to facts regarding to incidents, most of the incidents happened due to multi vehicle collision followed by single vehicle collision. Claims regarding incident of parked car and vehicle theft are very less. The collision in most of the incidents are from rear side followed by side collision. And in majority of incidents minor damage is reported followed by total loss. After the incident the authority contacted by the customer is mainly the police followed by fire authorities. Ambulance and other services were also contacted in many incidents. Most of the incidents have happened in New York (NY) and South Carolina (SC) and the cities which tops the incidents rate are Springfield, Arlington and Columbus. According to the dataset the most of the incidents have happened in mid night around 1am or 3am or in the evening at around 5pm. Single vehicle is involved in majority of claims followed by 3 vehicle and maximum vehicles involved are 4. There is at least one witness in most of the cases, there are cases in which no witness was present. In most of the claims police reports were not made available by the customer. Vehicles which are insured are mainly manufactured by Saab, Dodge, Suburu and Nissan. Most of the vehicles are manufactured in 1995 and the latest vehicles insured in the dataset were manufactured in 2015.

Inspecting Independent variables in claims where fraud was reported: -

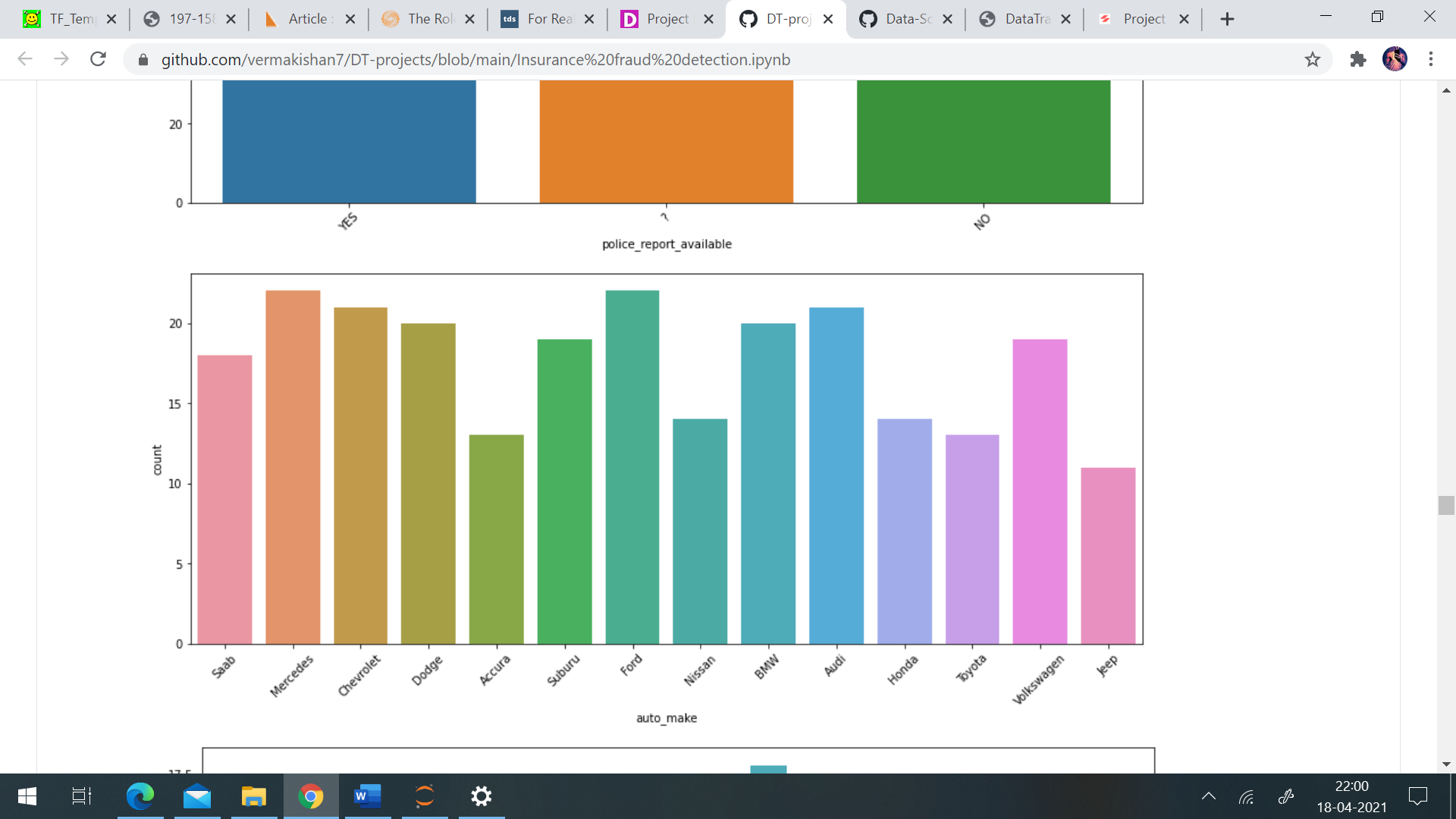
The majority of customers who have reported false claims were around 41 years old and most of them were female, it shows that young generation is not much involved in fraud claims. These customers do have a high education level, most of them have JD and MD. Customers having occupation of ecex-managerial are more likely to claim fraud insurance. One interesting thing to note down is that the customers who have done fraud claims have a hobby of playing chess. Fraud claims included most of the rear collision which occurred due to multi-vehicle or single vehicle collision. Major damage in reported in majority of fraud claims as it will result in claiming more money. Another interesting thing to notice is that in most of the fraud cases the authorities contacted was not police or fire or ambulance, other authorities were contacted in most of these cases.







Most of the incidents which were fraud are from SC state and least are from PA. The city in which most of fraud incidents were reported is Arlington followed by Columbus. Fraud insurance were reported mainly at around 4pm. Number of vehicles involved in these claims is mainly a single vehicle followed by 3 vehicles. As the insurance claim is fraud therefore, property damage is unknown in most of the cases. Bodily injuries are 2 in most of the fraud cases followed by 0 i.e., no bodily injuries. Fraud claims do have witnesses, majority of them have 2 witnesses followed by 1 witness. Police reports were not made available in most of the cases by the customers during claiming the fraud incident. Majority of fraud claims were reported of auto make Mercedes, Ford and Audi, as these are expensive vehicles. So the amount of claim will be also high for these vehicles.



Policy annual premium ranges from 433 to 2047.59 and the average premium is 1256.40. The total claim amount ranges from 100 to 1,14,920. The maximum injury claim is 21,450, maximum property claim is 23,670 and maximum vehicle claim is 79,560.

Correlation among the variables: -

I’ve plotted a heatmap for checking the correlation among the variables. Month as customer and age had a correlation of 0.92, means they have a very strong correlation. Total claim amount is strongly correlated with injury claim, property claim and vehicle claim, as these three claims contribute towards the total claim amount.

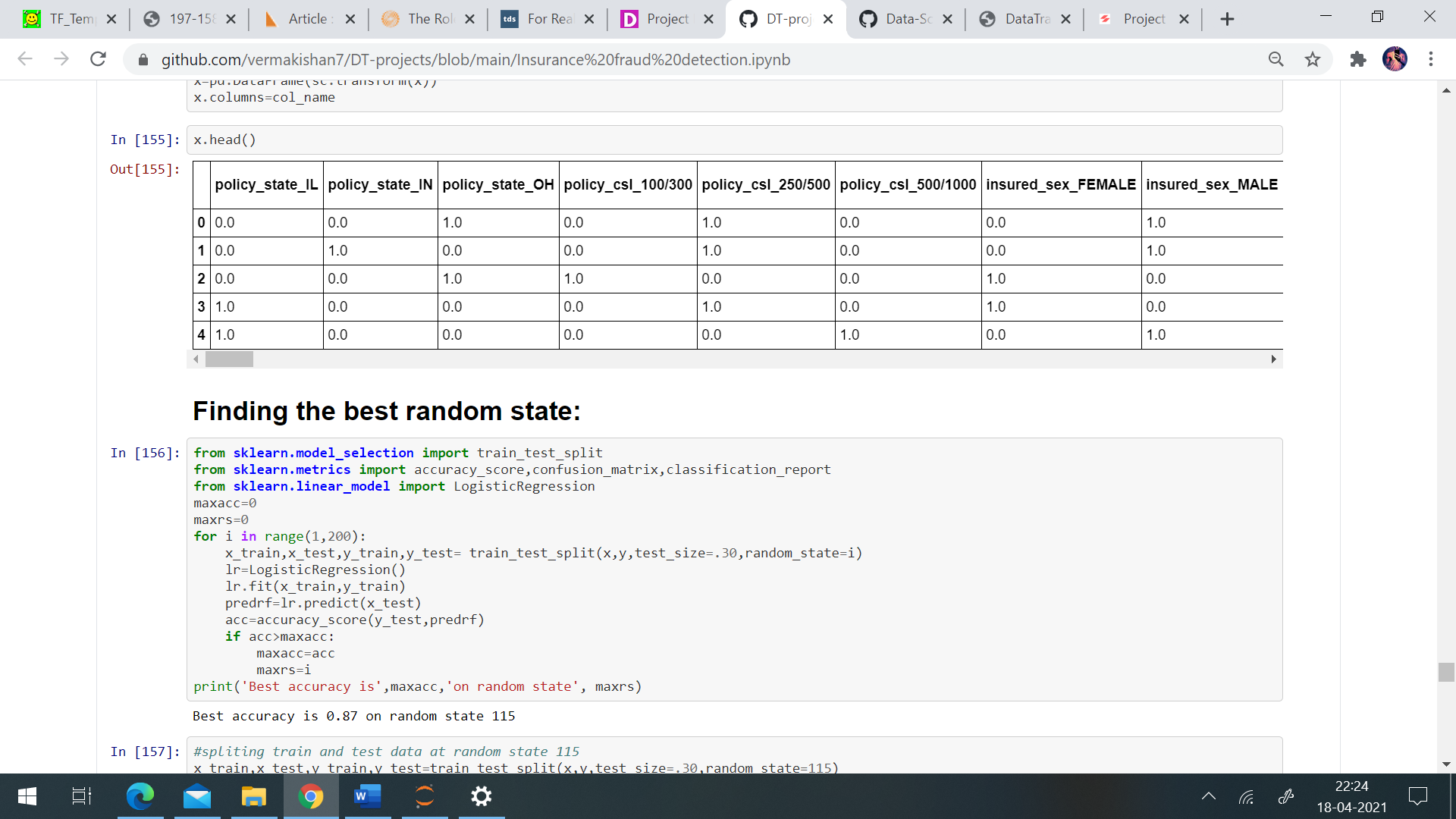
Checking the continuous variables: -

Most of the customers are customers of the insurance company from 100-250 months. Policy annual premium have a normal distribution. Capital gains and capital loss data is skewed. Total claim amount and vehicle claim have bimodal distribution. Injury claim and property claim data is skewed.

After this all the categorical columns were converted into numerical columns using one hot encoding, as machine learning models only takes numerical data. I have used label encoder to convert collision type into numerical column and have manually replaces ‘YES’ with 1, ‘NO’ and ‘?’ with 0 in property damage column and police report available column. In the same way I’ve manually converted the target variable into numerical by replacing ‘Y’ with 1 and ‘N’ with 0.

**Building Machine Learning Model:**

Before building the ML model, first separate the independent variable and dependent variable as x and y respectively. After that I’ve applied MinMaxScaler on the independent variables to bring them to the same scale (between 0 and 1). I’ve started by finding the best random state for the dataset, I’ll be doing the train test split at the random state on which highest accuracy is obtained. I’ve obtained an accuracy of 87% on random state 115. Now I’ve done train test split at random state 115.



I’ve used the following classification models:

LogisticRegression

GaussianNB

RandomForestClassifier

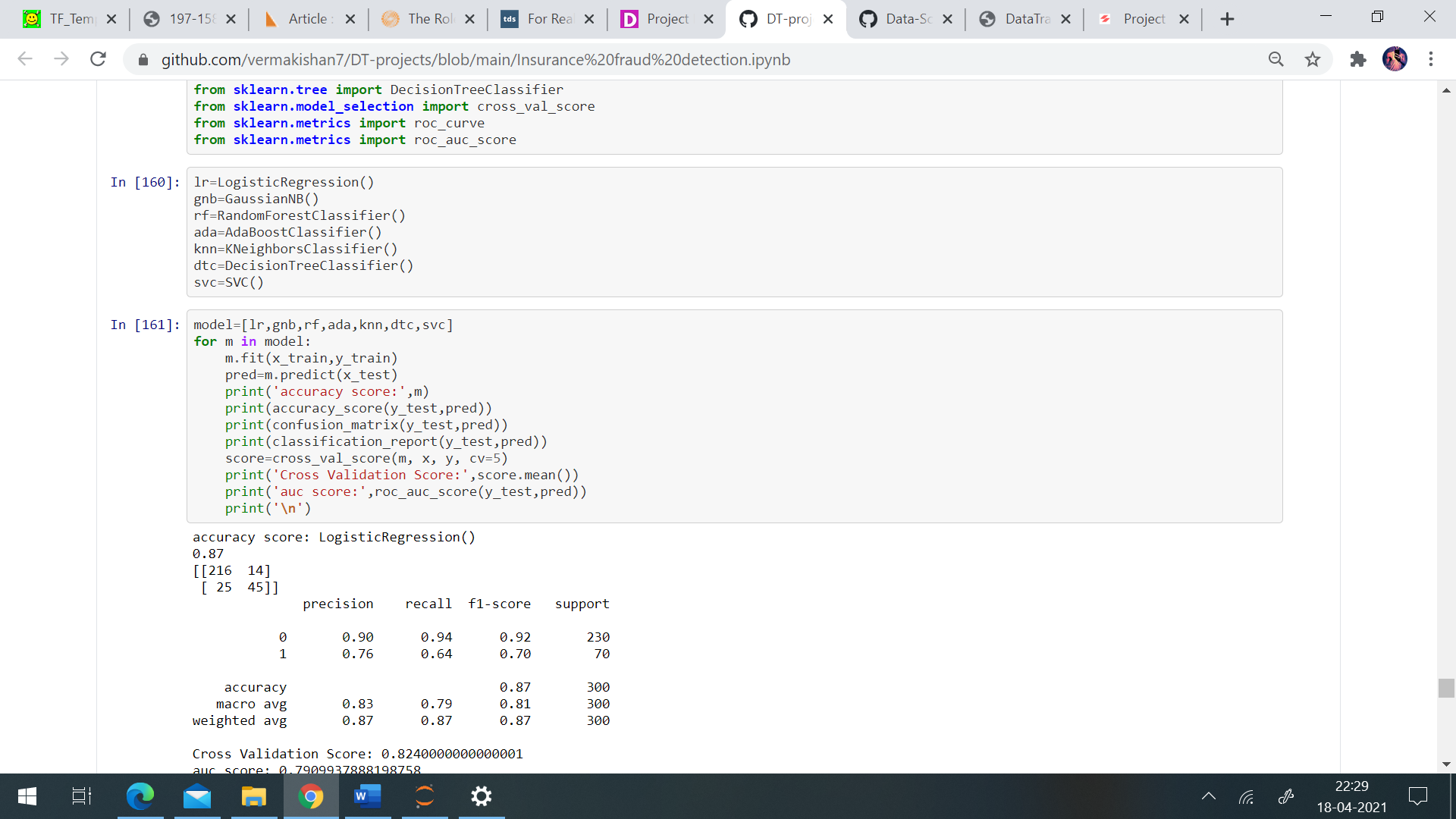
AdaBoostClassifier

KNeighborsClassifier

DecisionTreeClassifier

SVC

After importing the above models from respected libraries, I’ve trained the models using a for loop.



The result of the following loop is :

accuracy score: LogisticRegression()

0.87

[[216 14]

[ 25 45]]

precision recall f1-score support

0 0.90 0.94 0.92 230

1 0.76 0.64 0.70 70

accuracy 0.87 300

macro avg 0.83 0.79 0.81 300

weighted avg 0.87 0.87 0.87 300

Cross Validation Score: 0.8240000000000001

auc score: 0.7909937888198758

accuracy score: GaussianNB()

0.67

[[148 82]

[ 17 53]]

precision recall f1-score support

0 0.90 0.64 0.75 230

1 0.39 0.76 0.52 70

accuracy 0.67 300

macro avg 0.64 0.70 0.63 300

weighted avg 0.78 0.67 0.70 300

Cross Validation Score: 0.61

auc score: 0.7003105590062112

accuracy score: RandomForestClassifier()

0.8

[[219 11]

[ 49 21]]

precision recall f1-score support

0 0.82 0.95 0.88 230

1 0.66 0.30 0.41 70

accuracy 0.80 300

macro avg 0.74 0.63 0.65 300

weighted avg 0.78 0.80 0.77 300

Cross Validation Score: 0.7849999999999999

auc score: 0.6260869565217392

accuracy score: AdaBoostClassifier()

0.8133333333333334

[[211 19]

[ 37 33]]

precision recall f1-score support

0 0.85 0.92 0.88 230

1 0.63 0.47 0.54 70

accuracy 0.81 300

macro avg 0.74 0.69 0.71 300

weighted avg 0.80 0.81 0.80 300

Cross Validation Score: 0.798

auc score: 0.6944099378881988

accuracy score: KNeighborsClassifier()

0.7366666666666667

[[204 26]

[ 53 17]]

precision recall f1-score support

0 0.79 0.89 0.84 230

1 0.40 0.24 0.30 70

accuracy 0.74 300

macro avg 0.59 0.56 0.57 300

weighted avg 0.70 0.74 0.71 300

Cross Validation Score: 0.741

auc score: 0.5649068322981367

accuracy score: DecisionTreeClassifier()

0.76

[[194 36]

[ 36 34]]

precision recall f1-score support

0 0.84 0.84 0.84 230

1 0.49 0.49 0.49 70

accuracy 0.76 300

macro avg 0.66 0.66 0.66 300

weighted avg 0.76 0.76 0.76 300

Cross Validation Score: 0.792

auc score: 0.6645962732919255

accuracy score: SVC()

0.8266666666666667

[[210 20]

[ 32 38]]

precision recall f1-score support

0 0.87 0.91 0.89 230

1 0.66 0.54 0.59 70

accuracy 0.83 300

macro avg 0.76 0.73 0.74 300

weighted avg 0.82 0.83 0.82 300

Cross Validation Score: 0.8119999999999999

auc score: 0.7279503105590062

Logistic regression performed best for our data set with accuracy of 87%, cross validation score of 82%, auc score of 79% and f1-score of 92 and 70.

Then I’ve performed hyper parameter tuning for the best model i.e., Logistic regression.

The results after hyper parameter tunning are:

accuracy score:

0.87

[[216 14]

[ 25 45]]

precision recall f1-score support

0 0.90 0.94 0.92 230

1 0.76 0.64 0.70 70

accuracy 0.87 300

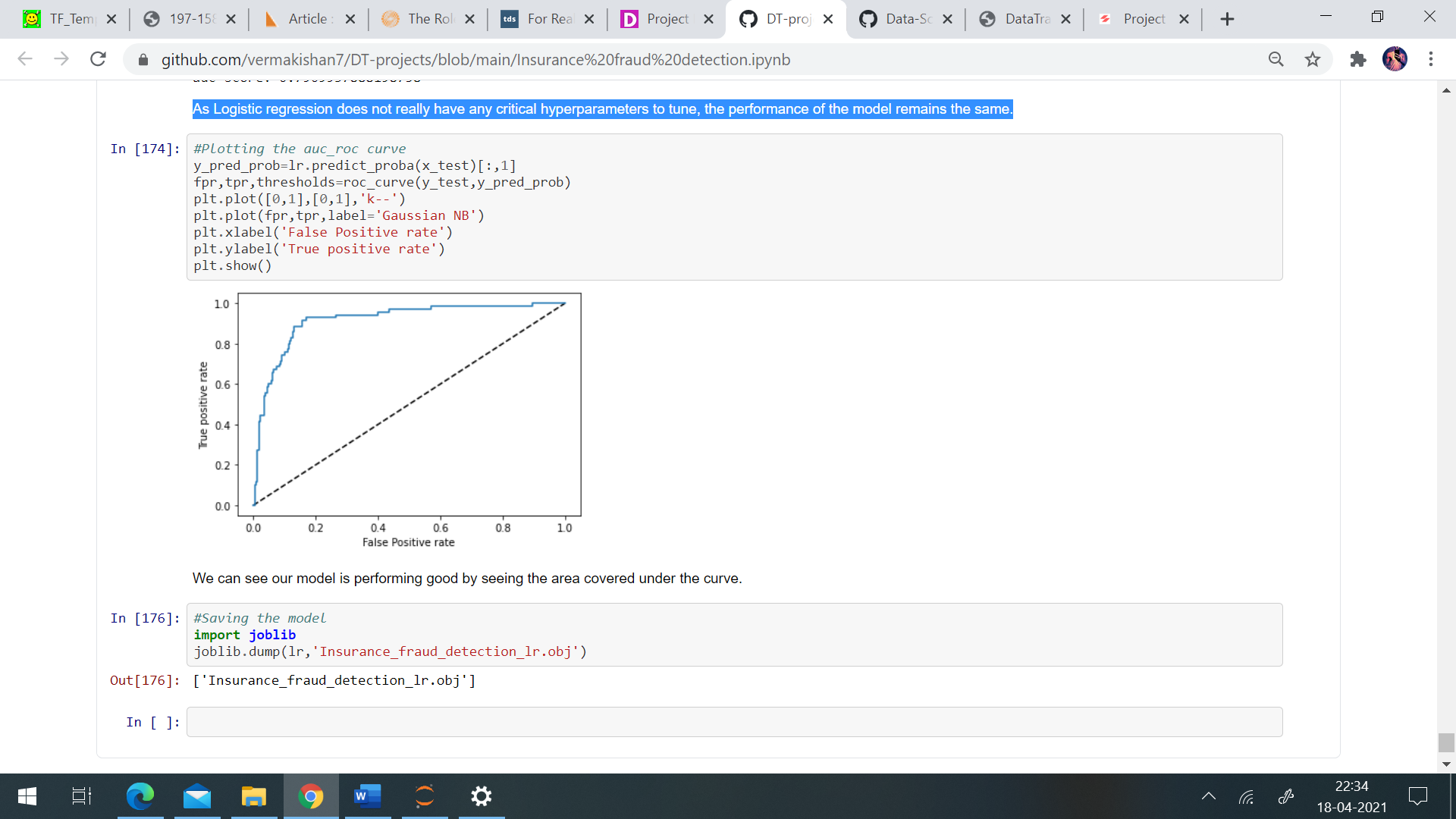
macro avg 0.83 0.79 0.81 300

weighted avg 0.87 0.87 0.87 300

Cross Validation Score: 0.8240000000000001

auc score: 0.7909937888198758

As Logistic regression does not really have any critical hyperparameters to tune, the performance of the model remains the same. Now lets see our model performance by plotting auc roc curve i.e., the curve between False positive rate and True positive rate.



**Conclusion:**

This project has built a model that can detect auto insurance fraud. In doing so, the model can reduce loses for insurance companies. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims.

I’ve trained 7 different classification models in this project. The best fitted model have an accuracy score of 87%, auc roc score of 79%, cross validation score of 82% and F-1 score of 0.70.