## 

## FRAUDULENT TRANSACTIONS SEGMENTATION USING UNSUPERVISED LEARNING

## CAPSTONE PROJECT REPORT

## Group – 7

## PGP DSE Gurgaon Jan 2020

Submitted towards fulfillment of the criteria for PGP-DSE Certification by GLIM

**Project Mentor & Research Supervisor:** Mr. RAM KUMAR

## SUBMITTED BY:

Alka Agarwal

Vanshika Arora

Nikhil Choudhary

Tushar Jethani

Kunal Jha

# **Acknowledgement**

We wish to place on record our deep appreciation for the guidance and help provided to us by our Mentor Mr. Ram Kumar. He helped us narrow down on the choice of the Project as well as the scope and focus area of the Project. He gave us valuable feedback at every stage to enhance the process and the outputs.

We would also like to place on record our appreciation for the guidance provided by Ms. Akshita Sawhney, Mrs. Namrata Thukral and Mr. Sachin Arya for giving us valuable feedback and being a source of inspiration in helping us to work on this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: 25/09/2020 Vanshika Arora

Place: Gurgaon Alka Agarwal

Nikhil Choudhary

Tushar Jethani

Kunal Jha

# 

# 

# **Certificate of Completion**

I hereby certify that the project titled “**FRAUDULENT TRANSACTIONS SEGMENTATION USING UNSUPERVISED LEARNING**” was undertaken and completed under my supervision by Alka Agarwal, Vanshika Arora, Nikhil Choudhary, Tushar Jethani & Kunal Jha, students of the Postgraduate Program in Data Science & Engineering (PGP DSE-JANUARY2020).

Date: September 25, 2020 (Ram Kumar)

Place: Gurgaon Mentor

# **TABLE OF CONTENTS**

# PAGE

# ACKNOWLEDGEMENTS ….....................................................................………2 CERTIFICATE OF COMPLETION .......................................................................... 3

# TABLE OF CONTENTS .........................................................................…4

# 1. Industry Review...................................................………………………

1.1 Industry Review – Current practices, Background Research

1.2 Literature Survey - Publications, Application, past and undergoing research

# 2. Dataset and Domain...........................................………………………

2.1 Data Dictionary

2.2 Variable categorization (count of numeric and categorical)

2.3 Pre-Processing Data Analysis (count of missing/ null values, redundant columns, etc.)

2.4 Alternate sources of data that can supplement the core dataset (at least 2-3 columns)

2.5 Project Justification - Project Statement, Complexity involved, Project Outcome –

Commercial, Academic or Social value

# 3. Data Exploration (EDA)................................................................

3.1 Relationship between variables

3.1.1 Distribution of variables

3.1.2 Presence of outliers and its treatment

3.1.3 Statistical significance of variables

3.1.4 Multi- collinearity

# 4. Feature Engineering................................................................

4.1 Whether any transformations required

4.2 Scaling the data

4.3 Feature selection

4.4 Dimensionality reduction

# 5. Model Building ............................................................................

5.1 Clustering

5.2 Classification

# 6. Conclusion …………………………………………………………………………………………..

# 7. Business Recommendations and Future Enhancements………………………….

7.1 How to improve data collection, processing and model accuracy?

7.2 Business Recommendations

1. **INDUSTRY REVIEW**

**INTRODUCTION**

Frauds using cell-phones, insurance-claims, tax-return claims, credit-card transactions, government procurement etc. represent significant problems for governments and businesses and specialized analysis techniques for discovering fraud using them are required. These methods exist in the areas of Knowledge Discovery in Databases (KDD), Data Mining, Machine Learning and Statistics. They offer applicable and successful solutions in different areas of electronic fraud crimes.

Despite the introduction of more secure chip technology, credit card fraud is still one of the biggest concerns of banks and credit card companies. After all, fraudsters continually change their tactics to get around any security measures that are put in place. A recent Nilson report projects that global total loss due to credit card scams and fraud will be a whopping $32.96 billion by 2021. Many financial companies are now looking at predictive analytics for fraud detection to see if they can depress these numbers.

The fraud threat facing banks and payments firms has grown dramatically in recent years. Estimates of fraud’s impact on consumers and financial institutions vary significantly but losses to banks alone are conservatively estimated to exceed $31 billion globally by 2018. Several converging trends have propelled the increasing scale, diversity, and complexity of fraud. Vulnerabilities in payments services have increased as the shift to digital and mobile customer platforms accelerates. New solutions have also led to payments transactions being executed more quickly, leaving banks and processors with less time to identify, counteract, and recover the underlying funds when necessary. Finally, the sophistication of fraud has increased, in part through greater collaboration among bad actors, including the exchange of stolen data, new techniques, and expertise on the dark web.

Increasingly agile fraud perpetrators have benefited from banks’ and payments firms’ limited ability to adapt. While most institutions have well-funded anti-fraud groups, key resources are often fragmented across the organization. Essential data, investigative and forensics expertise, and analytics talent are typically distributed across cyber, compliance, legal, IT, and fraud teams, with little to no coordination or data sharing.

Effectively combating fraud through analytics requires a mindset shift from a narrow focus on false positives and loss prevention to an appreciation that the same technological advancements making fraud more pervasive also enable the tools and environment to address it. With their shift to digital services, banks have access to exponentially more customer and transaction data than in the past. New technologies create the means to more accurately segment customers by risk, enabling lower-friction digital experiences—and higher satisfaction levels—for low-risk customers. And the explosion of industry verticals in cyber and data analytics has created a ready supply of talented, cross-disciplinary resources unencumbered by legacy organizational structures. Today’s challenge is harnessing these components to reduce current losses, detect and prevent emerging fraud, and enhance customer experience.

# **1.1 Current Practices**

In order to effectively test, detect, validate, correct error and monitor control systems against fraudulent activities, businesses entities and organizations rely on specialized data analytics techniques such as data mining, data matching, sounds like function, Regression analysis, Clustering analysis and Gap. Techniques used for fraud detection fall into two primary classes: statistical techniques and artificial intelligence.

In general, the primary reason to use data analytics techniques is to tackle fraud since many internal control systems have serious weaknesses. For example, the currently prevailing approach employed by many law enforcement agencies to detect companies involved in potential cases of fraud consists in receiving circumstantial evidence or complaints from whistleblowers. As a result, a large number of fraud cases remain undetected and unprosecuted.

Analytics approaches to detect & prevent Frauds:

* Combine historical fraud data with industry knowledge & external market data
* Create a proof of concept to test the history data to determine fraud cases
* If historical data is not available, then anomaly detection or outlier detection is used
* Apply the statistical model for fraud detection
* Models are based on past spending patterns, demographic information
* Further text mining & link analysis for probable associations to find deeper frauds

Early data analysis techniques were oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations and can help to get better insights into the processes behind the data. Although the traditional data analysis techniques can indirectly lead us to knowledge, it is still created by human analysts.

To go beyond, a data analysis system has to be equipped with a substantial amount of background knowledge and be able to perform reasoning tasks involving that knowledge and the data provided. In effort to meet this goal, researchers have turned to ideas from the machine learning field. This is a natural source of ideas, since the machine learning task can be described as turning background knowledge and examples (input) into knowledge (output).

If data mining results in discovering meaningful patterns, data turns into information. Information or patterns that are novel, valid and potentially useful are not merely information, but knowledge. One speaks of discovering knowledge, before hidden in the huge amount of data, but now revealed.

The machine learning and artificial intelligence solutions may be classified into two categories: 'supervised' and 'unsupervised' learning. These methods seek for accounts, customers, suppliers, etc. that behave 'unusually' in order to output suspicion scores, rules or visual anomalies, depending on the method.

Whether supervised or unsupervised methods are used, note that the output gives us only an indication of fraud likelihood. No standalone statistical analysis can assure that a particular object is a fraudulent one, but they can identify them with very high degrees of accuracy.

# **Supervised learning**

In supervised learning, a random sub-sample of all records is taken and manually classified as either 'fraudulent' or 'non-fraudulent' (task can be decomposed on more classes to meet algorithm requirements). Relatively rare events such as fraud may need to be over sampled to get a big enough sample size. These manually classified records are then used to train a supervised machine learning algorithm. After building a model using this training data, the algorithm should be able to classify new records as either fraudulent or non-fraudulent.

Supervised neural networks, fuzzy neural nets, and combinations of neural nets and rules, have been extensively explored and used for detecting fraud in mobile phone networks and financial statement fraud.

Bayesian learning neural network is implemented for credit card fraud detection, telecommunications fraud, auto claim fraud detection, and medical insurance fraud.

Hybrid knowledge/statistical-based systems, where expert knowledge is integrated with statistical power, use a series of data mining techniques for the purpose of detecting cellular clone fraud. Specifically, a rule-learning program to uncover indicators of fraudulent behavior from a large database of customer transactions is implemented.

Cahill et al. (2000) design a fraud signature, based on data of fraudulent calls, to detect telecommunications fraud. For scoring a call for fraud its probability under the account signature is compared to its probability under a fraud signature. The fraud signature is updated sequentially, enabling event-driven fraud detection.

Link analysis comprehends a different approach. It relates known fraudsters to other individuals, using record linkage and social network methods.

This type of detection is only able to detect frauds similar to those which have occurred previously and been classified by a human. To detect a novel type of fraud may require the use of an unsupervised machine learning algorithm.

# **Unsupervised learning**

In contrast, unsupervised methods don't make use of labelled records.

Some important studies with unsupervised learning with respect to fraud detection should be mentioned. For example, Bolton and Hand use Peer Group Analysis and Break Point Analysis applied on spending behavior in credit card accounts. Peer Group Analysis detects individual objects that begin to behave in a way different from objects to which they had previously been similar. Another tool Bolton and Hand develop for behavioral fraud detection is Break Point Analysis. Unlike Peer Group Analysis, Break Point Analysis operates on the account level. A break point is an observation where anomalous behavior for a particular account is detected. Both the tools are applied on spending behavior in credit card accounts.

Artificial Intelligence has been at work for quite a while at differing levels in bank and credit card companies. AI development has only gotten better since online transactions started to turn up the volume. Many of us would have received a telephone call, a text or email alert when an unusual transaction was detected. This was the first time that pattern recognition software came to the forefront. The usual general rules of fraudulent behavior were set aside and it was found to be much more effective to create patterns for each account. Pattern recognition software or neural networks have looked at patterns in individual accounts and cards to know when something seems out of tune. Automated notifications are sent out to the customer and, in many cases, transactions are stopped.

In recent times, companies have turned to using data analytics to detect fraud. Predictive analytics has expanded the capabilities for fraudulent transaction detection and has experienced a wide adoption among banks and credit card companies in particular.

How companies are using it:

* Financial institutions using it to identify frauds in leasing contracts
* Banks are using it to detect credit card, wire transfers, check frauds
* Insurers are using it to detect fraudulent claims to save the losses
* Healthcare provider can optimize the medical loss ratio by detecting claims frauds

The estimation models have been built by researchers using ginormous data sets. Think 900 million transactions from about 7 million individual cards. Out of this, about 120,000 were known as fraudulent transactions. These researchers have used subsets of this data to test their model. Patterns of transactions, as mentioned earlier, continue to be what the predictive models are based on.  In the machine learning period of training, different highly customized variables are created based on looking ‘deeply’ at each and every transaction.

This can be more easily explained with a real-life transaction history. A customer is found to make 2 or 3 online purchases every month, and the amount spent never exceeds $500. This looks like a person who loves shopping and discount offers, so a variable will be created for this category. The machine then creates an if/then decision tree with features that will or will not point to fraud. Now, when a real-time transaction is run through this decision tree, the predictive model will decide then and there whether the transaction is fraudulent or not.

# **1.2 Literature Survey - Publications, Application, past and undergoing research**

1. **Machine Learning for Unsupervised Fraud Detection. (n.d.). RÉMI DOMINGUES**. <https://www.diva-portal.org/smash/get/diva2:897808/FULLTEXT01.pdf>

Fraud is a threat that most online service providers must address in the development of their systems to ensure an efficient security policy and the integrity of their revenue. Amadeus, a Global Distribution System providing a transaction platform for flight booking by travel agents, is targeted by fraud attempts that could lead to revenue losses and indemnifications. The objective of this thesis is to detect fraud attempts by applying machine learning algorithms to bookings represented by Passenger Name Record history. Due to the lack of labelled data, the current study presents a benchmark of unsupervised algorithms and aggregation methods. It also describes anomaly detection techniques which can be applied to self-organizing maps and hierarchical clustering. Considering the important amount of transactions per second processed by Amadeus back-ends, we eventually highlight potential bottlenecks and alternatives.

1. **(PDF) Combining unsupervised and supervised learning in credit card fraud detection. (n.d.).** <https://www.researchgate.net/publication/333143698_Combining_Unsupervised_and_Supervised_Learning_in_Credit_Card_Fraud_Detection>

Supervised learning techniques are widely employed in credit card fraud detection, as they make use of the assumption that fraudulent patterns can be learned from an analysis of past transactions. The task becomes challenging, however, when it has to take account of changes in customer behavior and fraudsters’ ability to invent novel fraud patterns. In this context, unsupervised learning techniques can help the fraud detection systems to find anomalies. In this paper we present a hybrid technique that combines supervised and unsupervised techniques to improve the fraud detection accuracy. Unsupervised outlier scores, computed at different levels of granularity, are compared and tested on a real, annotated, credit card fraud detection dataset. Experimental results show that the combination is efficient and does indeed improve the accuracy of the detection.

1. **(PDF) Combining unsupervised and supervised learning in credit card fraud detection. (n.d.).** <https://www.researchgate.net/publication/333143698_Combining_Unsupervised_and_Supervised_Learning_in_Credit_Card_Fraud_Detection>

[Machine learning has been instrumental in solving some of the important business problems](https://marutitech.com/problems-solved-machine-learning/) such as detecting email spam, focused product recommendation, accurate medical diagnosis etc. The adoption of machine learning (ML) has been accelerated with increasing processing power, availability of big data and advancements in statistical modeling. Fraud management has been painful for banking and commerce industry. The number of transactions has increased due to a plethora of payment channels – credit/debit cards, smartphones, kiosks. At the same time, criminals have become adept at finding loopholes. As a result, it’s getting tough for businesses to authenticate transactions. Data scientists have been successful in solving this problem with machine learning and predictive analytics. Automated fraud screening systems powered by machine learning can help businesses in reducing fraud.

1. **Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. (n.d.). Apapan Pumsirirat, Liu Yan**. <https://thesai.org/Downloads/Volume9No1/Paper_3-Credit_Card_Fraud_Detection_Using_Deep_Learning.pdf>

Frauds have no constant patterns. They always change their behavior; so, we need to use an unsupervised learning. Fraudsters learn about new technology that allows them to execute frauds through online transactions. Fraudsters assume the regular behavior of consumers, and fraud patterns change fast. So, fraud detection systems need to detect online transactions by using unsupervised learning, because some fraudsters commit frauds once through online mediums and then switch to other techniques. This paper aims to 1) focus on fraud cases that cannot be detected based on previous history or supervised learning, 2) create a model of deep Auto-encoder and restricted Boltzmann machine (RBM) that can reconstruct normal transactions to find anomalies from normal patterns. The proposed deep learning based on auto-encoder (AE) is an unsupervised learning algorithm that applies backpropagation by setting the inputs equal to the outputs. The RBM has two layers, the input layer (visible) and hidden layer. In this research, we use the TensorFlow library from Google to implement AE, RBM, and H2O by using deep learning. The results show the mean squared error, root mean squared error, and area under curve.

1. **Auto insurance fraud detection using unsupervised spectral ranking for anomaly. (n.d.).** <https://www.sciencedirect.com/science/article/pii/S2405918816300058>

For many [data mining](https://www.sciencedirect.com/topics/economics-econometrics-and-finance/data-mining) problems, obtaining labels is costly and time consuming, if not practically infeasible. In addition, unlabeled data often includes categorical or ordinal features which, compared with numerical features, can present additional challenges. We propose a new unsupervised spectral ranking method for anomaly (SRA). We illustrate that the spectral optimization in SRA can be viewed as a relaxation of an unsupervised SVM problem. We demonstrate that the first non-principal eigenvector of a Laplacian matrix is linked to a bi-class classification strength measure which can be used to rank anomalies. Using the first non-principal eigenvector of the Laplacian matrix directly, the proposed SRA generates an anomaly ranking either with respect to the majority class or with respect to two main patterns. The choice of the ranking reference can be made based on whether the cardinality of the smaller class (positive or negative) is sufficiently large. Using an auto insurance claim data set but ignoring labels when generating ranking, we show that our proposed SRA significantly surpasses existing outlier-based fraud detection methods. Finally, we demonstrate that, while proposed SRA yields good performance for a few similarity measures for the auto insurance claim data, notably ones based on the Hamming distance, choosing appropriate similarity measures for a fraud detection problem remains crucial.

1. **Neurospace. (2019, March 29). Predicting credit card fraud with unsupervised learning.** <https://neurospace.io/blog/2019/03/predicting-credit-card-fraud-with-unsupervised-learning/>

This report shows how well an unsupervised learning algorithm can detect fraudulent transactions, when it has been trained only on normal transactions. The hypothesis is, that when having learned the patterns of normal transactions, we can take the model to production and start using machine learning algorithms for detecting fraudulent behavior earlier than if we have to wait for several frauds to happen, so we can label them as fraud and then later use supervised learning algorithms.

1. **Streaming Active Learning Strategies for Real-Life Credit Card Fraud Detection: Assessment and Visualization. (n.d.). Fabrizio Carcillo · Yann-A¨el Le Borgne · Olivier Caelen · Gianluca Bontempi**. <https://arxiv.org/pdf/1804.07481.pdf>

Credit card fraud detection is a very challenging problem because of the specific nature of transaction data and the labeling process. The transaction data is peculiar because they are obtained in a streaming fashion, they are strongly imbalanced and prone to non-stationarity. The labeling is the outcome of an active learning process, as every day human investigators contact only a small number of cardholders (associated to the riskiest transactions) and obtain the class (fraud or genuine) of the related transactions. An adequate selection of the set of cardholders is therefore crucial for an efficient fraud detection process.

1. **Credit-Card Fraud Profiling Using a Hybrid Incremental Clustering Methodology. (n.d.). Marie-Jeanne Lesot, Adrien Revault d’Allonnes**. <https://hal.archives-ouvertes.fr/hal-01282307/document>

This paper addresses the task of helping investigators identify characteristics in credit-card frauds, so as to establish fraud profiles. To do this, a clustering methodology based on the combination of an incremental variant of the linearized fuzzy c-medoids and a hierarchical clustering is proposed. This algorithm can process very large sets of heterogeneous data, i.e. described by both categorical and numeric features. The relevance of the proposed approach is illustrated on a real dataset containing next to one million fraudulent transactions.

1. **An application of unsupervised fraud detection to Passenger Name Records. (n.d.). Remi Domingues, Francesco Buonora, Romain Senesi, Olivier Thonnard**. <https://www.eurecom.fr/en/publication/5058/download/data-publi-5058.pdf>

Fraud is a threat that most online service providers must address in the development of their systems to ensure an efficient security policy and the integrity of their revenue. If rule-based systems and supervised methods usually provide the best detection and prevention, labelled training datasets are often non-existent and such solutions lack reactivity when facing adaptive fraudsters. Many generic fraud detection solutions have been made available for companies though cannot compete with dedicated internal implementations.

# **2. DATA SET AND DOMAIN**

# **2.1 Data Dictionary**

**Variable information/Data description**

|  |  |
| --- | --- |
| **NAME** | **DESCRIPTION** |
| FLOW  FEATURE | Flow feature include identifying features which describe network traffic flow between connected packets of hosts. |

|  |  |
| --- | --- |
| proto | Transaction protocol. A protocol is a set of rules that govern how the communication should take place.  A protocol is a standard used to define a method of exchanging data over a computer network, such as local area network, Internet, Intranet, etc. Each protocol has its own method of how to handle data in the following situations:   * How data is formatted when sent. * What to do with data once received. * How data is compressed. * How to check for errors in the data. |

|  |  |
| --- | --- |
| BASIC FEATURES | Basic features include features which represent basic protocol header information. |
| state | The last-known or current status of an application or a process. The Internet is intrinsically stateless because each request for a new Web page is processed without any knowledge of previous pages requested.  Indicates to the state and its dependent protocol, e.g. ACC, CLO, CON, ECO, ECR, FIN, INT, MAS, PAR, REQ, RST, TST, TXD, URH, URN, and (-) (if not used state). It will used as an application to analyse the collected packets to:   * detect anomalous network traffic which could indicate a potential security problem with a computer or a violation of the department's Conditions of Use, * detect attacks against our web servers and potentially malicious attacks from our clients against remote servers |
| dur | Record total duration of a transactions in seconds |
| sbytes | The total number of bytes used in transaction from the source protocol to the destination protocol |
| dbytes | The total number of bytes used in transaction from the destination protocol to the source protocol |
| sttl | Source to destination time to live value. Time to live (TTL) refers to the amount of time or “hops” that a packet is set to exist inside a network before being discarded by a router. |
| dttl | Destination to source time to live value |
| sloss | Source packets retransmitted or dropped. When accessing the internet or any network, small units of data called packets are sent and received. When one or more of these packets fails to reach its intended destination, this is called packet loss |
| dloss | Destination packets retransmitted or dropped |
| service | A networking service is a low-level application that enables the network to perform more than basic functions.  http, ftp, smtp, ssh, dns, ftp-data ,irc and (-) if not much used service. Different types of service are used for different purpose likewise, smtp is simple mail transfer protocol used as a communication protocol for electronic transmissions. In the same way, FTP is file transfer protocol helps in transferring files from source to destination. |
| Sload | Source bits per second. The data rate is a term to denote the transmission speed, or the number of bits per second transferred. The useful data rate for the user is usually less than the actual data rate transported on the network. One reason for this is that additional bits are transferred for e.g signalling, the address, the recovery of timing information at the receiver or error correction to compensate for possible transmission errors.  In telecommunications, it is common use to express the data rate in bits per seconds (bit/s), see bitrate. In datacommunication, the data rate is often expressed in bytes per second (B/s). |
| Dload | Destination bits per second |
| Spkts | Source to destination packet count. It is the hop count refers to the number of intermediate network devices through which data must pass between source and destination. Hop count is a rough measure of distance between two hosts. |
| Dpkts | Destination to source packet count |
| CONTENT  FEATURES | Content features include attributes of TCP header information pertaining to content size, source IP address and packet flow. |
| swin | Source TCP window advertisement value. A **TCP window advertisement** determines the maximum amount of data that can be sent before the sender must wait for an acknowledgement from the receiver. By **advertising** its **window** size, the receiver side manages flow control. |
| dwin | Destination TCP window advertisement value |
| stcpb | Source TCP base sequence number. All bytes in a **TCP** connection are numbered, beginning at a randomly chosen initial **sequence number** (ISN). The SYN packets consume one **sequence number**, so actual data will begin at ISN+1. The **sequence number** is the byte **number** of the first byte of data in the **TCP** packet sent (also called a **TCP** segment). |
| dtcpb | Destination TCP base sequence number |
| smeans | Mean of the packet size transmitted by the src. **SRC** is Series Session and Resource Control Modules give service providers a dynamic **network** resource allocation solution that enables them to deliver differentiated products and services. |
| dmeans | Mean of the packet size transmitted by the dst. Distribution spanning tree is A dynamic ad hoc **network** consists of a collection of mobile hosts with frequently changing **network** topology. |
| trans\_depth | Represents the pipelined depth into the connection of http request/response transaction. |
| res\_bdy\_len | Actual uncompressed content size of the data transferred from the server’s http service. |
| TIME  FEATURES | Time features include features of time, for example, inter‐arrival time between packets start/end packet time and round trip time of TCP protocol. |
| Sjit | Source jitter (mSec). jitter can refer to packet delay variation, the variation (statistical dispersion) in the delay of the packets. |
| Djit | Destination jitter (mSec) |
| Stime | record start time |
| Ltime | record last time |
| Sintpkt | Source interpacket arrival time (mSec).  It is a measure of the times between packets arriving at a host over a period.  If packets are arriving at a host every second, the mean inter-packet arrival time is 1 second. Of course, network traffic doesn't happen that neatly. So depending on what you want to know, you would measure the times between packet arrival (usually a relative time) and calculate the mean. |
| Dintpkt | Destination interpacket arrival time (mSec) |
| tcprtt | TCP connection setup round-trip time, the sum of ’synack’ and ’ackdat’. |
| synack | TCP connection setup time, the time between the SYN and the SYN\_ACK packets. SYN is used to initiate and establish a connection. It also helps you to synchronize sequence numbers between devices.  SYN\_ACK- SYN message from local device and ACK of the earlier packet. |
| ackdat | TCP connection setup time, the time between the SYN\_ACK and the ACK packets.  ACK Helps to confirm to the other side that it has received the SYN. |
| is\_sm\_ips\_ports | If source and destination IP addresses are equal and port numbers are equal then, this variable takes value 1 else 0 |
| ADDITIONAL GENERATED  FEATURES | Additionally generated features include general purpose features which each feature has its own purpose, in order to protect the service of protocols and connection features are built from the flow of 100 record connections based on the sequential order of the last time feature. |
| ct\_state\_ttl | No. for each state according to specific range of values for source/destination time to live. |
| ct\_flw\_http\_mthd | No. of flows that has methods such as Get and Post in http service. |
| is\_ftp\_login | If the ftp session is accessed by user and password then 1 else 0. |
| ct\_ftp\_cmd | No of flows that has a command in ftp session. |
| ct\_srv\_src | No. of connections that contain the same service and source address in 100 connections according to the last time. |
| ct\_srv\_dst | No. of connections that contain the same service and destination address in 100 connections according to the last time. |
| ct\_dst\_ltm | No. of connections of the same destination addres in 100 connections according to the last time. |
| ct\_src\_ ltm | No. of connections of the same source address in 100 connections according to the last time. |
| ct\_src\_dport\_ltm | No of connections of the same source address and the destination port in 100 connections according to the last time. |
| ct\_dst\_sport\_ltm | No of connections of the same destination address and the source port in 100 connections according to the last time. |
| ct\_dst\_src\_ltm | No of connections of the same source and the destination address in in 100 connections according to the last time. |
| TARGET FEATURES | |
| attack\_cat | The name of each attack category. In this data set, nine categories e.g. Fuzzers, Analysis, Backdoors, DoS Exploits, Generic, Reconnaissance, Shellcode and Worms |
| Label | 0 for normal and 1 for attack records |

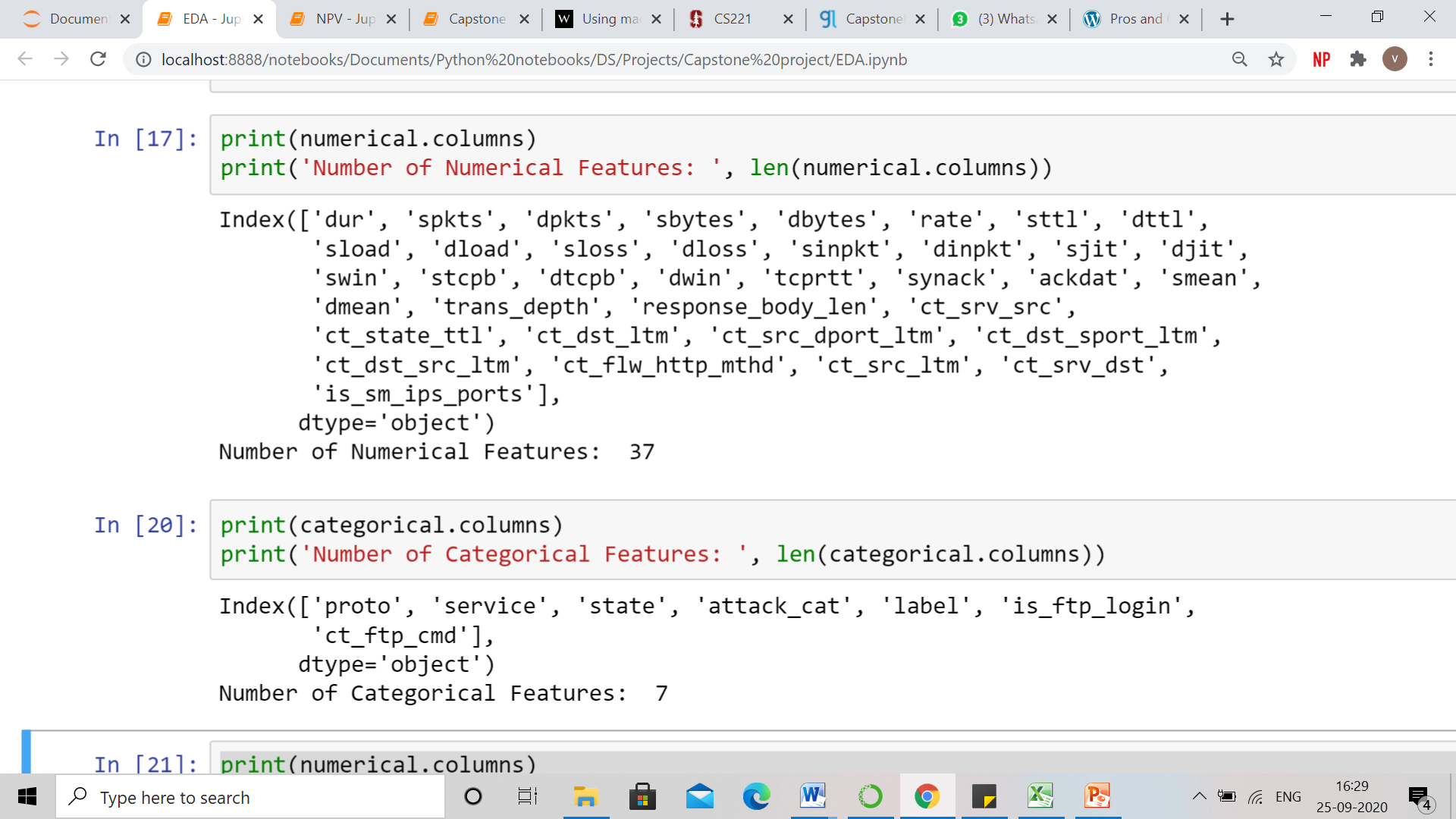
# **2.2 Variable Categorization**

The dataset has got both numerical as well as categorical features. However, some features have been assigned as numeric datatype by python but these features have only few limited and fixed possible values.

Defining the categorical and numerical features

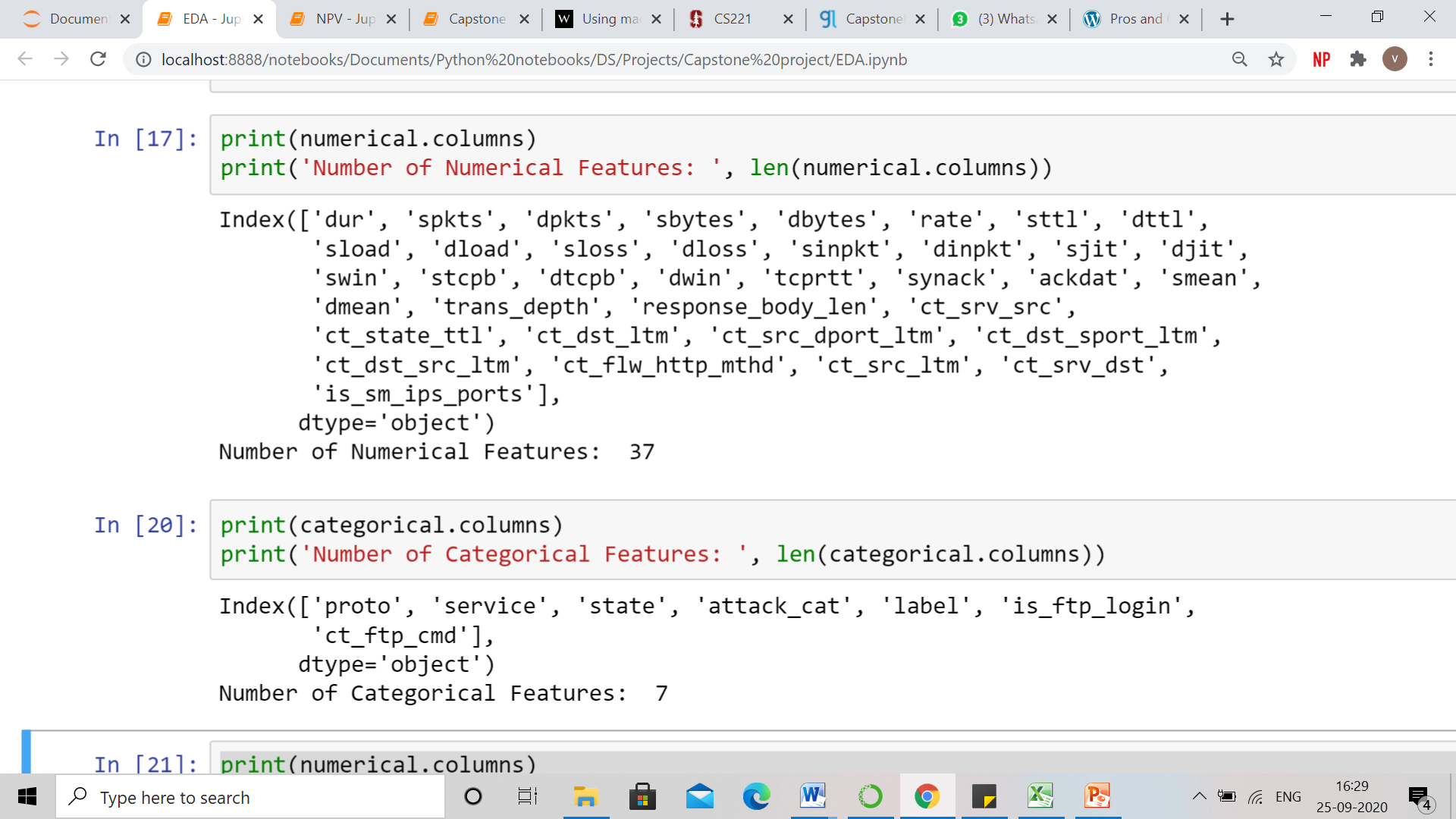
**Numerical Features:-**

There are total 37 numerical features including the target feature of label.

****

**Categorical Features:-**

There are total 7 categorical features including the target feature of attack category.



# **2.3 Pre Processing Data Analysis (count of missing/ null values, redundant columns)**

**Missing Values:-**

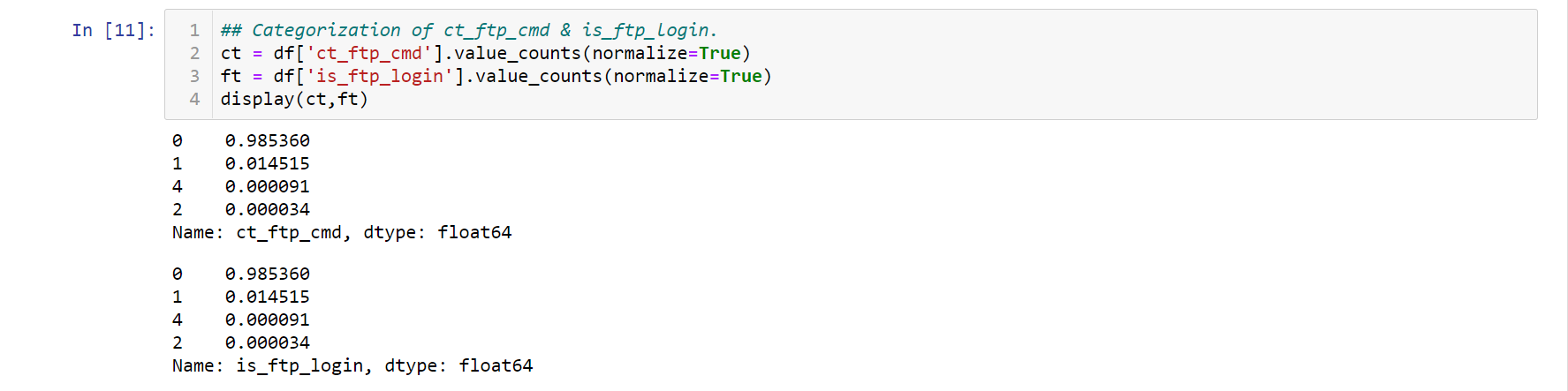
There are no null or missing values in the dataset (df). Here, is the snapshot of our Jupiter notebook.



**Redundant Columns:-**

**Here are few of the redundant columns along with their reasons.**

* Id: - Id column contains all unique values/ serial number of each row which is of no use in our analysis. Hence we have dropped Id column.
* Tcprtt : It is basically the sum of 2 columns i.e. synack and ackdat. In order to remove redundancy from the data, we have removed this column.
* ct\_ftp\_cmd : Majority of data of this column is similar to the data of is\_ftp\_login column. Hence, in order to avoid multicollinearity, we have removed this column. Below is the snapshot of the same.



**Target** **Columns**: -

While doing Unsupervised learning for segmenting fraudulent transactions, there is no target column (As we are making clusters/groups of similar records and not predicting anything). However, while doing Classification (Supervised learning), we have taken the below mentioned columns as target columns.

NOTE: We have also performed several statistical tests to find out whether the features are significant for predicting the target feature.

**Labels: -**

It is a categorical feature which is comprised of 2 categories.

These categories are:

**0 represents “No attack happened”**

**1 represents “Attack happened”**

**Attack Category: -**

It is also a categorical feature which is comprised of 9 attack type categories and 10th is normal or no attack. All attack type categories are mentioned in detail below:-

There are nine attacks types discovered in the Dataset.

* ***Fuzzers*** : An attack in which the attacker tries to discover security loopholes in the Operating System, program or network and make these resources suspended for some time period and can even crash them.
* ***Analysis***: A type intrusions that penetrate the web applications through port scanning, malicious web scripting and dispatching spam emails etc.
* ***Backdoor***: A technique in which attacker can bypass the usual authentication and can get unauthorized remote access to a system.
* ***DoS***: An intrusion in which attacker tries to disrupt the computing resources, by making them extremely busy in order to prevent the authorized access to the resources.
* ***Exploit***: The intrusions which utilize the software vulnerabilities, error or glitch within the operating systems (OS) or software.
* ***Generic***: This attack act against a crypto-graphical system and it tries to break the key of the security system.
* ***Reconnaissance***: It can be defined as a probe; an attack that gathers information about the target computer network in order to bypass its security control.
* ***Shell code***: A malware attack in which the attacker penetrates a slight piece of code starting from a shell to control the compromised machine.
* ***Worm***: malware that replicate themselves and spread to other computers by using the network to spread the attack, depending on the security failures on the target computer which it wants to access.

# **2.4 Alternate sources of data that can supplement the core dataset (at least 2-3 columns)**

We were able to find out alternate dataset from the similar topic of network intrusion. We found it out on kaggle, the name of the dataset is “NSL\_KDD data”. In the dataset, we found 11 features which were namely, “dst\_bytes”, dst\_host\_srv\_diff\_host\_rate, srv\_diff\_host\_rate, land, dst\_host\_same\_src\_port\_rate, logged\_in, num\_root, srv\_rerror\_rate.

# **2.5 Project Justification - Project Statement, Complexity involved, Project Outcome – Commercial, Academic or Social value**

# **Introduction**

A network intrusion is any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources and data and jeopardizing the security of network and their data. With advancement in technology these network intrusions and fraudulent transactions have increased. These unauthorized activities almost always imperil the security of networks and their data.

Nowadays, online brands and companies are the usual subjects of these attacks. Information gathered from this study can be passed onto cyber security teams of such organizations and will help organizations have an in-depth understanding of how network intrusions and fraudulent transactions work and effect formidable detection and prevention systems.

# **Project Statement**

Using Unsupervised Learning, segmenting the attacks and provide an analysis report for the cyber security team to take action.

# **Complexity involved**

While applying unsupervised learning to form clusters so that appropriate attack types can be clubbed together and useful insights can be derived, here are a few challenges that we faced while working on this data:

* **Size of the Dataset:** The original dataset was divided into 4 parts and there were more than 15 lakhs rows in the data. It was not feasible for us to work on such huge data in such a short span of time. A separate sample train dataset was also available which consists of 1,75,341 rows and 45 columns. We chose the train dataset, however a different set of problems occurred on this dataset also which are mentioned in next few points.
* **Combining Quantitative and Qualitative data for clustering:** While most of the popular algorithms for clustering like K-means & Agglomerative clustering are designed in a way that they provide best results when they are applied on a numerical set of features. Not much research has happened on clustering algorithms which can be applied on numerical and categorical data.
* **Time constraint**: As the size of data is huge and the data is also a mixture of categorical and numerical features, running Top down clustering would take time as in Top down clustering all observations start in one cluster, and splits are performed recursively as one move down the hierarchy. Top down clustering is a strategy of hierarchical clustering. Also, other types of clustering algorithms fail as they are unable to process both (numerical and categorical) kinds of data smoothly.
* **Non-Applicability of K-mode Algorithm**: K-mode algorithm for clustering works on categorical data and it came out in [a paper of 1998 by Zhexue Huang](http://link.springer.com/article/10.1023/A:1009769707641) as an alternate of K-means. However, this algorithm fails to run on mixed data containing both numerical and categorical data. Also, there’s very little documentation about it.
* **Underdeveloped Algorithm like K-Prototype and Gower distance:** Not much research has happened on these algorithms which resulted in underdevelopment of these algorithms. These algorithms are highly slow when compared to it’s counterpart like K-means and that is the reason why these algorithms fails to run on a larger dataset.

# **Project Outcome – Commercial**

Organizations can’t protect themselves from every type of hacker or every form of intrusion.

Organizations of all sizes, in virtually every industry on earth, were hit by cyberattacks. From shipping giant Maersk, to the ride sharing giant Uber, all the way to Equifax, a credit rating agency regarded as one of the largest holders of private customer data in the world.

With advancement in technology, intrusions and cyber-attacks have increased, which is causing privacy issues, monetary loss, time loss for businesses. Of course, the bigger the company, the greater the cost of downtime caused by an attack. Considering that that average downtime per company was estimated to be 23 hours in 2017, the monetary cost of each inoperable hour could be devastating.

Cyberattacks on small businesses are more common than many think, with more than two-thirds (67 percent) of companies with fewer than 1,000 employees having experienced a cyberattack, and 58 percent having experienced a breach, according to a recent [report](https://keepersecurity.com/assets/pdf/Keeper-2018-Ponemon-Report.pdf).

The same report also found that 60 percent of small businesses could go out of business due to damages associated with a cyberattack.

Information like medical records, credit card information, Social Security numbers, bank account credentials or proprietary business information is then accessible. In 2019, for example, a small medical practice in Michigan reported a ransomware attack that had encrypted their files, including patient records, appointment schedules and payment information. The two doctors who owned the business refused to pay a ransom to unlock their files, according to a published [report](https://www.hipaajournal.com/michigan-practice-forced-to-close-following-ransomware-attack/). After the hackers deleted their files, the owners closed their doors.

Whether it is online or in the physical world, payment systems now operate in a digital ecosystem. This makes everyone using this ecosystem vulnerable to cyberattacks, both merchants and consumers.

All the above points showcase the necessity of network intrusion system that can detect fraudulent transactions and alert the organization regarding the same so that appropriate steps can be taken in development of more advance systems.

Here are few steps for businesses to tighten their network security:-

* **Secure your hardware -** With so much attention given to acquiring the newest and most sophisticated types of cyber security software, safeguarding the security of company hardware is often overlooked but the loss or theft of devices is a real threat to be aware of. Begin your cyber-attack prevention strategy with the basics: protect all devices with a complicated password, [share that password with the device user only](https://www.entrepreneur.com/article/292612) and commit it to memory instead of writing it down in an easily accessible place. Do not overlook the effectiveness of physically attaching computers to desks. This is a simple, yet effective way of preventing intruders from walking away with company equipment and the sensitive data they hold. Finally, install ‘find my device’ software on all laptops, phones and tablets. By doing so, equipment that is stolen can quickly be located by the authorities.
* **Encrypt and back up data** - An effective cyber-crime protection strategy must consist of two elements: preventing physical access to sensitive data and rendering that data useless if it falls into the wrong hands. Companies can achieve the latter by always encrypting their data. As highlighted by researchers in the International Journal of Advanced Computer Science and Applications, data encryption remains the ‘most efficient fix’ for data breaches, should they occur. Be sure to encrypt all sensitive data, including customer information, employee information and all business data. Full-disk encryption software is included in virtually all operating systems today and can encrypt all the data on a desktop or laptop computer when it’s at rest.
* **Stay ahead by backing up data and storing it separately**- After encryption, backing up all data is another key way of protecting yourself from security breaches. With ransomware hackers locking companies out of their systems, encrypting their data and asking for a ransom to be paid before releasing the data, you can stay one step ahead of them by backing up all of your data and storing it separately.
* **Educate staff on the dangers of unsecured networks**- Banning employees from using their personal devices for work may seem like an obvious approach, but this strategy seldom works in the long term. As staff members grow tired of the inconvenience, they are likely to return to accessing work on personal devices, regardless of policies prohibiting this.

# **Project Outcome – Academic or Social value**

Government agencies hold a treasure trove of confidential information, including fingerprints, Social Security numbers, and more. Government servers and databases, unfortunately, have known vulnerabilities, resulting in larger amounts and volumes of attacks in recent years. In 2016, a hacking group called the Shadow Brokers breached the NSA, highlighting the common and problematic practice of gathering intelligence through bugs in commercial products instead of notifying the software companies who make the software. That ill-advised practice can potentially endanger billions of software users.

Around the world, the digitization of government is gathering pace, with a host of interactions now carried out online. In some countries, you can vote, pay bills and taxes, and get medical prescriptions – often using a single, digital citizen ID that’s stored centrally.

This hasn’t escaped the attention of criminals that once focused primarily on retail banking and e-commerce. We’re seeing a rise in fraudulent personal and corporate tax and VAT returns and associated rebates, along with bogus welfare claims.

Data is leaking from both public and private sector organizations, either due to malicious hacking or rogue employees. Globally over 700 million personal data records were compromised in 2015, with the largest single breach exceeding information on 70 million individuals.

Cyber criminals are also getting better at ‘social engineering,’ in the form of subtle emails or phone calls from apparently legitimate sources such as banks, financial advisers or even lawyers. In some cases these emails are even sent from the IT systems of those trusted advisers once the cyber-criminal has broken into their email system – the so called business email compromise fraud.

Few things Government Agencies can do to minimise the risk:-

* **Treat cyber security as a critical organizational issue -** Security isn’t just something you can leave to the IT specialists. It affects everyone working in government. Take the oil and gas industry, where personal safety has long been paramount. Companies in this sector have tried to make cyber security an equally central part of their culture – alongside safety, and not just a ‘compliance’ issue. Employees are encouraged to think about what kinds of assets are at risk, and how they can prevent attacks and spot threats.
* **Encourage innovative cyber security solutions -**
  + If governments want to realize the savings and efficiencies from going digital, they need to constantly keep one step ahead of criminals.
  + Governments have to be even more nimble, to come up with innovative and cost-effective ways to block cyber-crime and frustrate the efforts of criminals to cash-out and monetize stolen information. New technologies such as biometrics, analytics and virtualization can play a part – but so can education and awareness.
  + Banks also adopt a philosophy known as ‘fast to fail,’ which halts unsuccessful projects quickly, before they consume too much money. By following this example, governments could become more agile, and develop systems that spot threats early and prevent breaches.
* **Collaborate with the private sector -**
  + Collaboration can bring in fresh, external thinking as well as providing challenge, benchmarking and peer comparisons. We bring our clients together to provide safe spaces for discussion, swapping war stories, and finding inspiration in each other’s experience. The global [I-4 conference programme](https://i4online.com/) is just one example of our work in this area.
  + Being prepared to share intelligence on actual and potential attacks also matters. After all, the kind of information floating around the criminal fraternity is often stolen from, and used against a combination of public and private organizations, so it’s in everybody’s interests to work together.
* **Embed more security into your supply chain –**
  + Today’s governments are often heavily dependent upon a wide and complex web of service providers and contractors. With so many parties processing confidential information, the chances for leaks or theft are much higher. The best way to counter this challenge is by tightening up procurement. Contracts should embed cyber security. Ideally suppliers should all be certified to an industry standard. Regular monitoring and independent audits can reassure government that standards are being maintained, to avoid weak links in the chain. Most importantly, make sure contracts drive the right behaviours when responding to a cyber security incident ensuring openness, transparency and a willingness to work together when the worst happens.
* **Plan your talent needs carefully -**
  + In future, governments should widen their collaboration with private companies to include talent sharing. Cyber security specialists could rotate roles between the public and private sectors, as part of their natural career development. It wouldn’t just help government; it would also give these individuals a higher personal profile.

# **3. EXPLORATORY DATA ANALYSIS (EDA)**

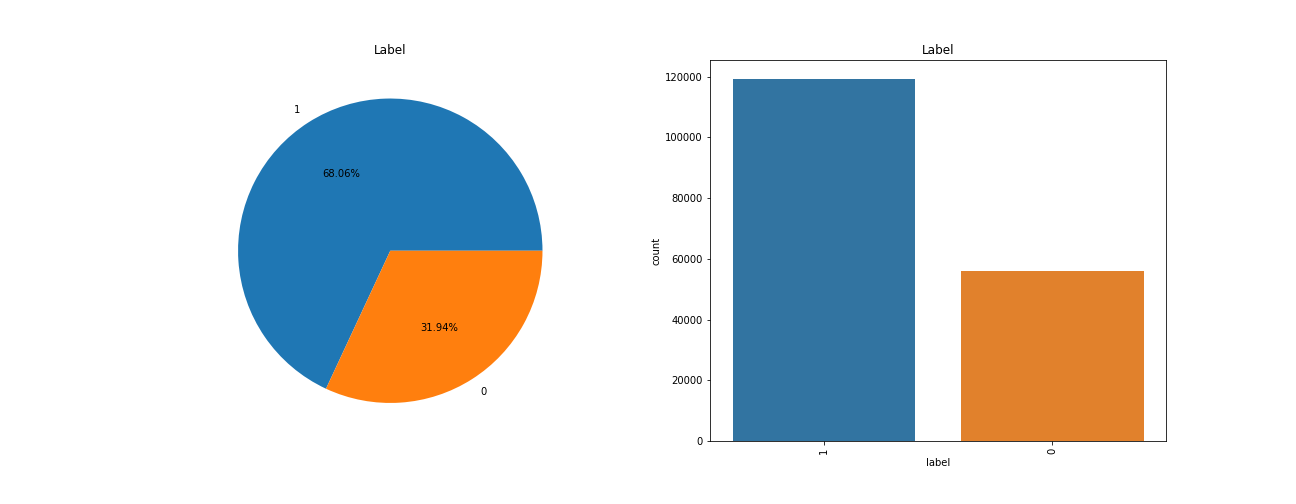
# **3.1 Relationship between variables**

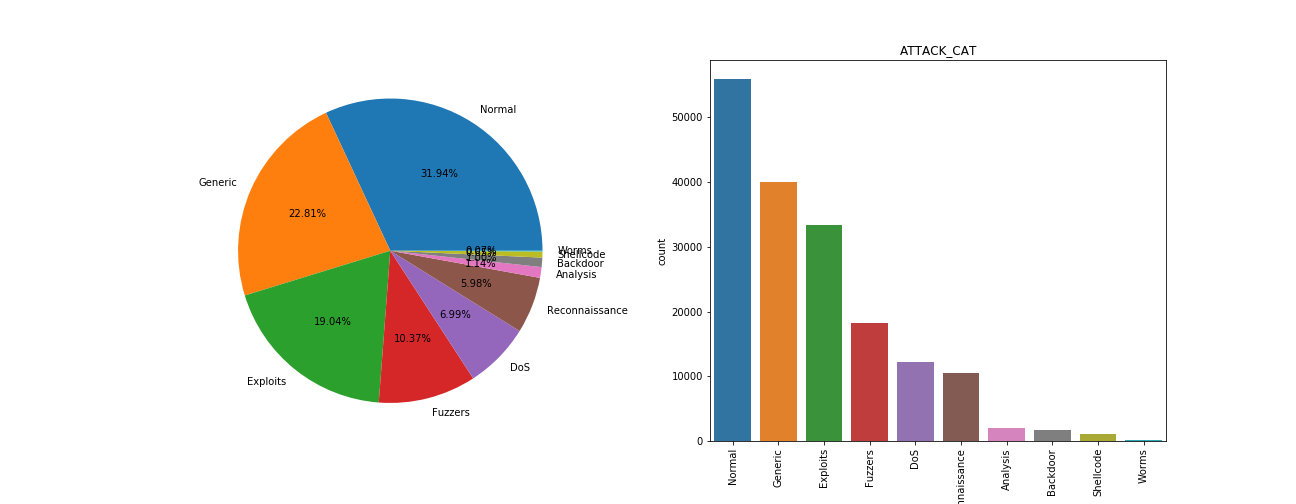
### 3.1.1 Distribution of variables

**Target Features**

Thereare 2 target features, first is label - which tells whether an attack has taken place (1) or not (0). Second is the attack category feature – it lists 9 different types of attacks and the 10th is normal or no attack.

68% of the labels are 1 which means signifies that an attack has taken place. Generic and Exploits are the common attacks whereas Analysis, Backdoor, Shellcode and Worms cyber attacks happen rarely.





There are 36 numerical features and 4 categorical features excluding the target variables and redundant variables.

**Categorical Features**

There are 5 categorical columns excluding the target features – proto, service and state, is\_ftp\_login and ct\_ftp\_cmd.

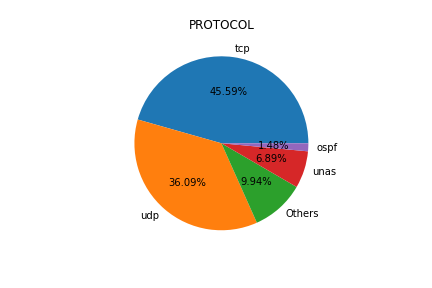
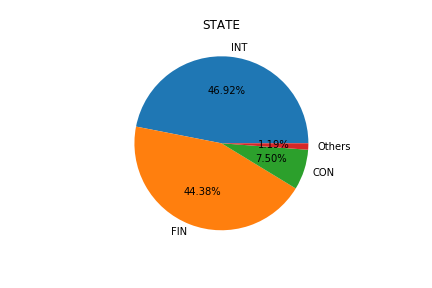
There are 3 major categories in proto – ‘tcp’, ‘udp’ and ‘unas’ which represents 88% of the data.

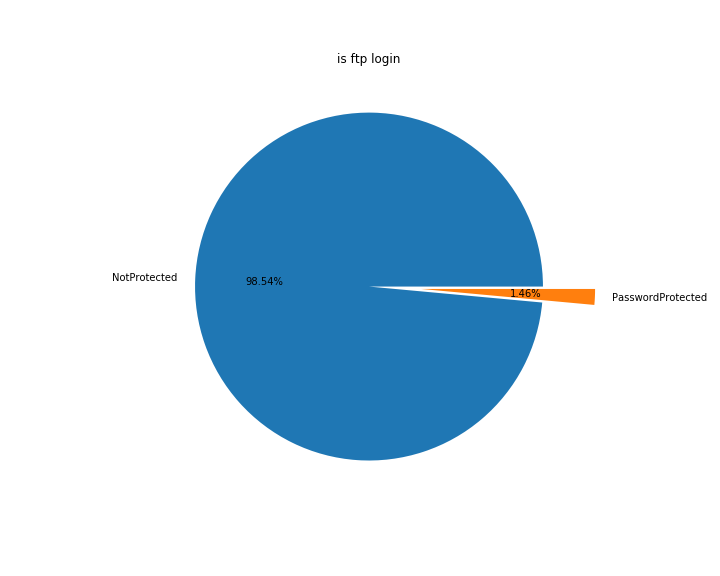
In service category, service ‘not used’, ‘dns’ and ‘http’ makes almost 91% of the service feature.

In state, ‘int’, ‘fin’ and ‘con’ makes almost 98 % of the state feature.

* Most of the transactions are done without using any service i.e. service=Not\_used.
* Most of the FTP sessions have been accessed without user and password.
* There are many categorical features which have numerous categories within them and their value count is very less. These sub-categories were binned together as ‘others’.

After, EDA it was observed that is ftp login had only 2 values and therefore, it was converted into a categorical column. 0 was replaced as ‘NotProtected’ and 1 was replaced as ‘PasswordProtected’. Most of the data is not protected by password.

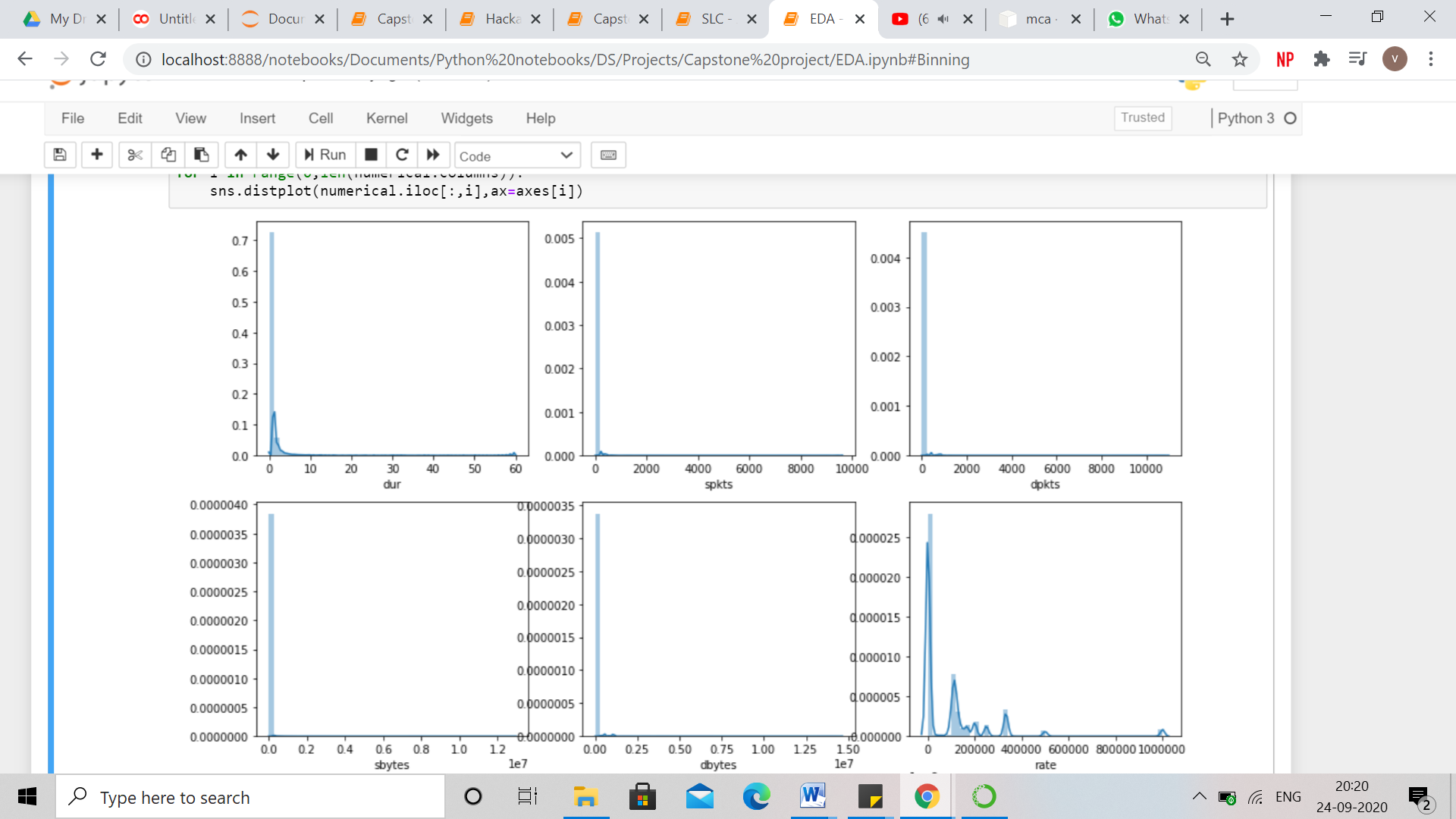
  

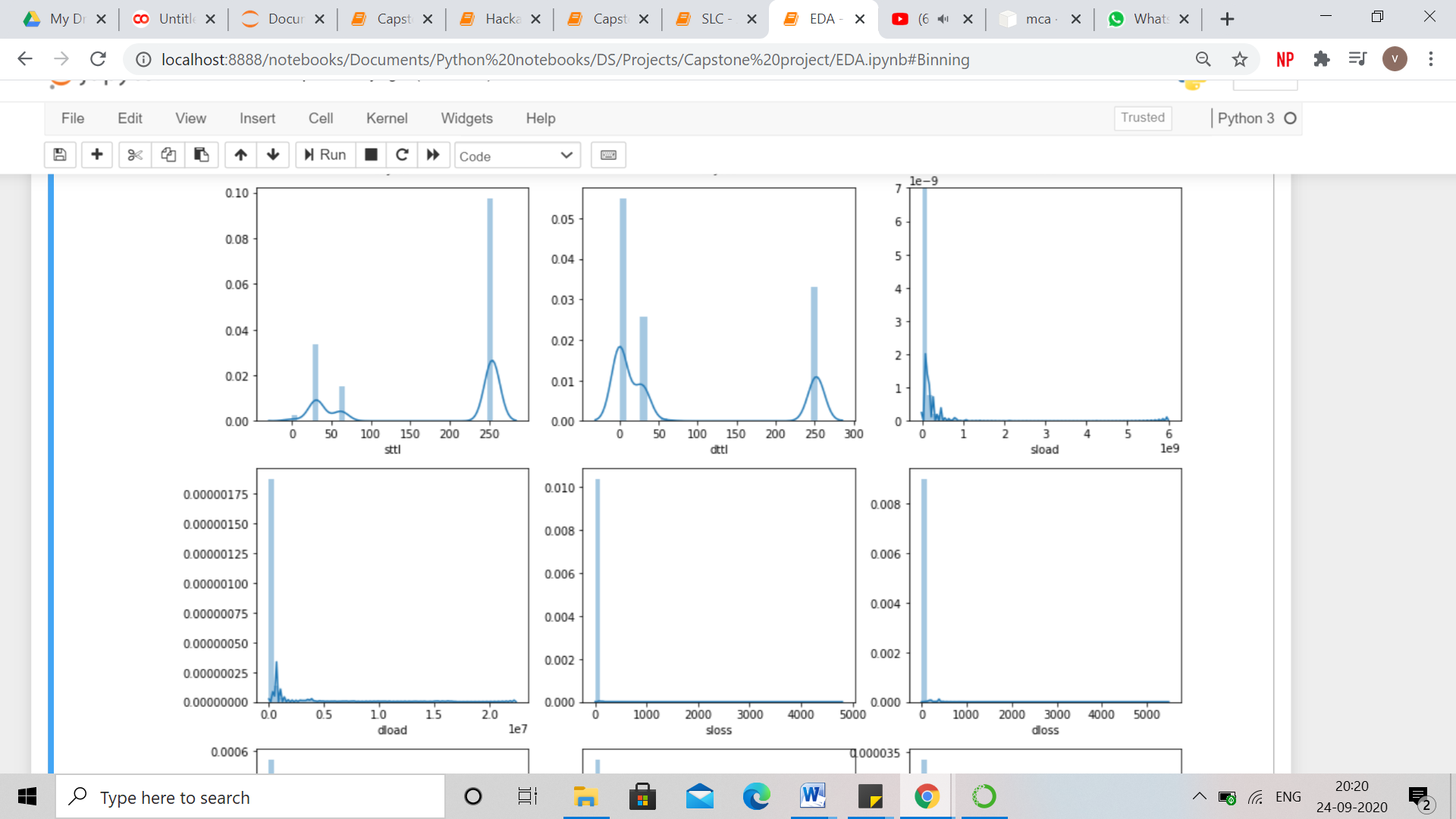
** **

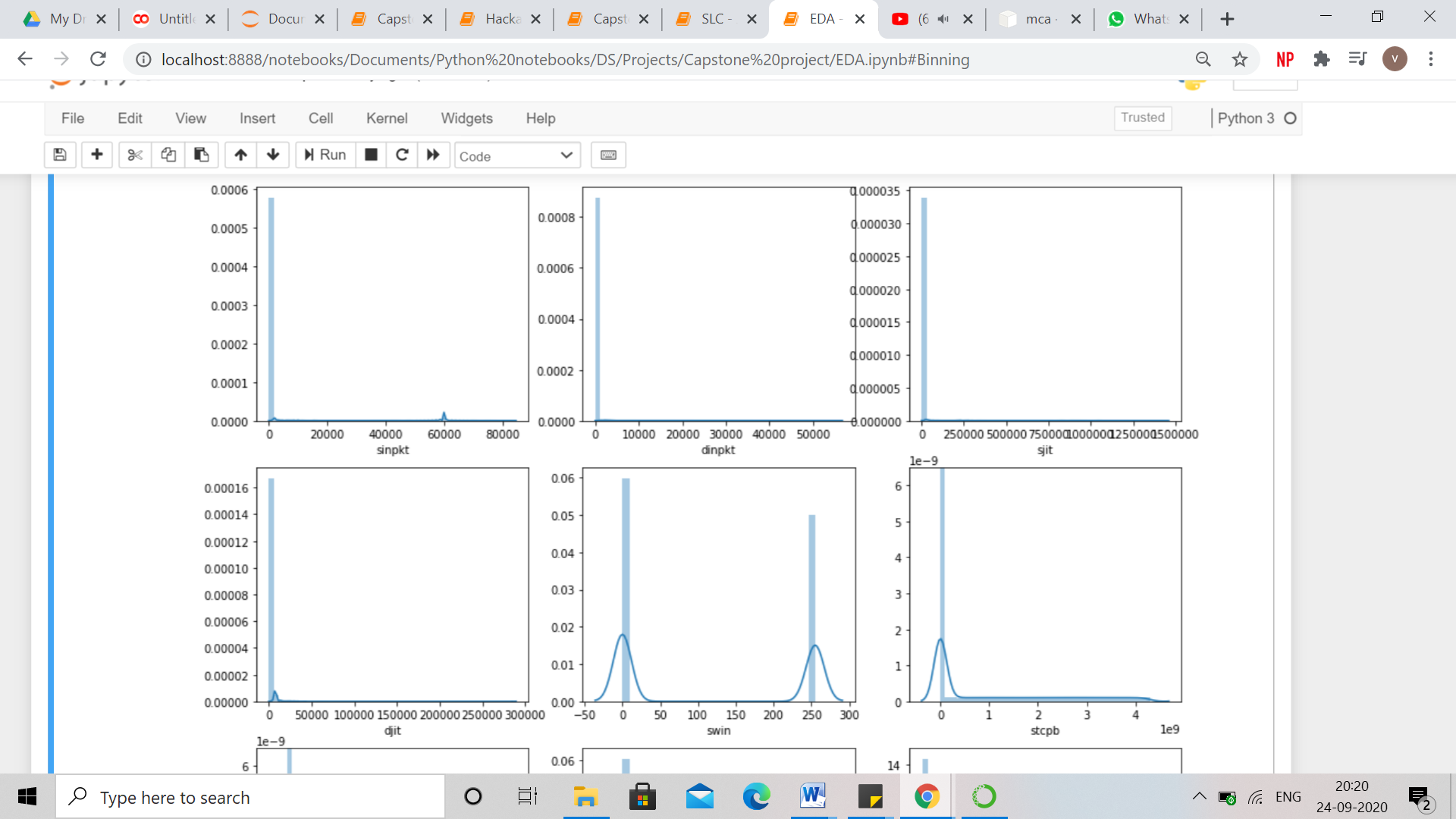
**Numerical Features**

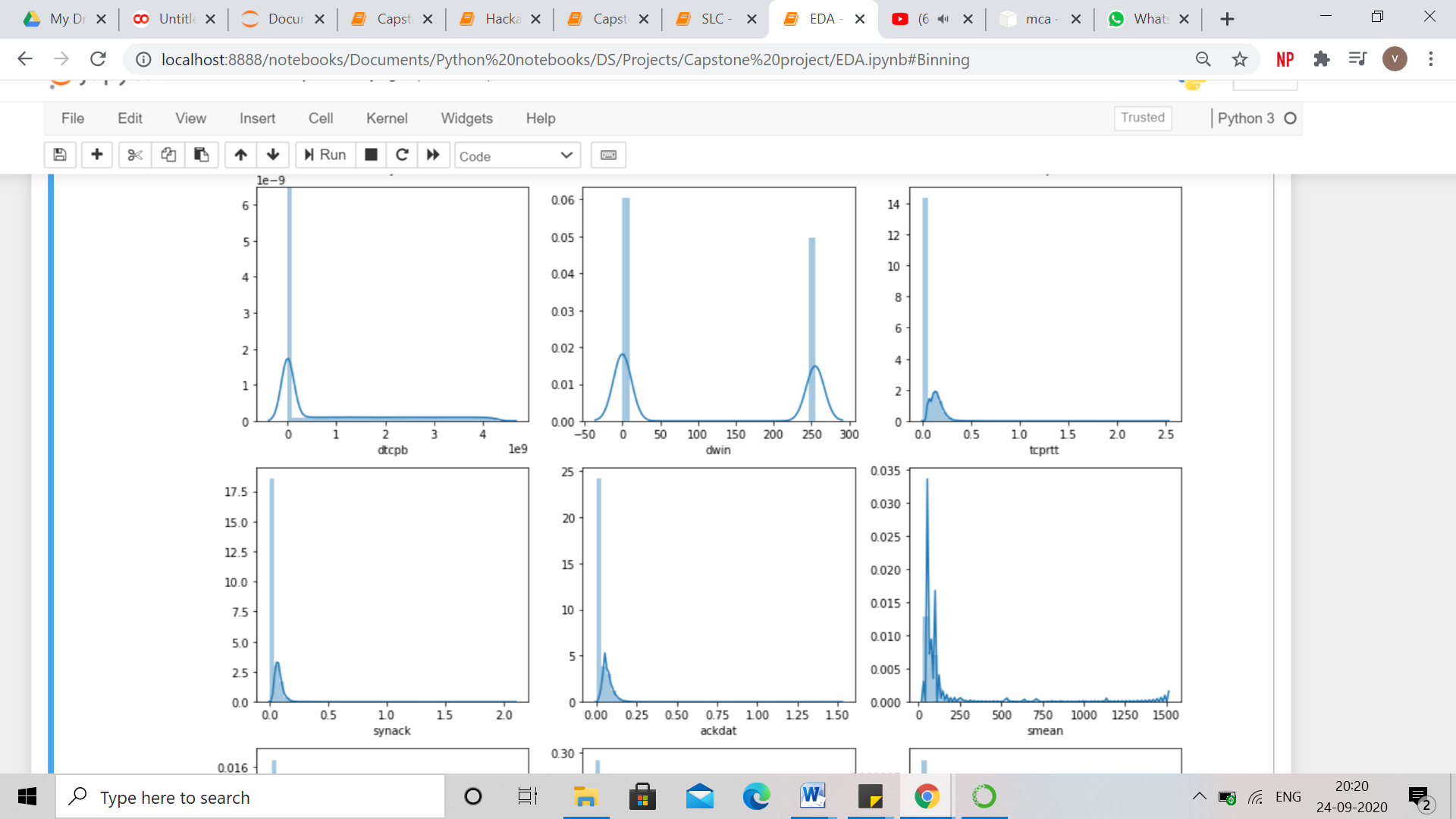
There are 37 numerical features present in the dataset.

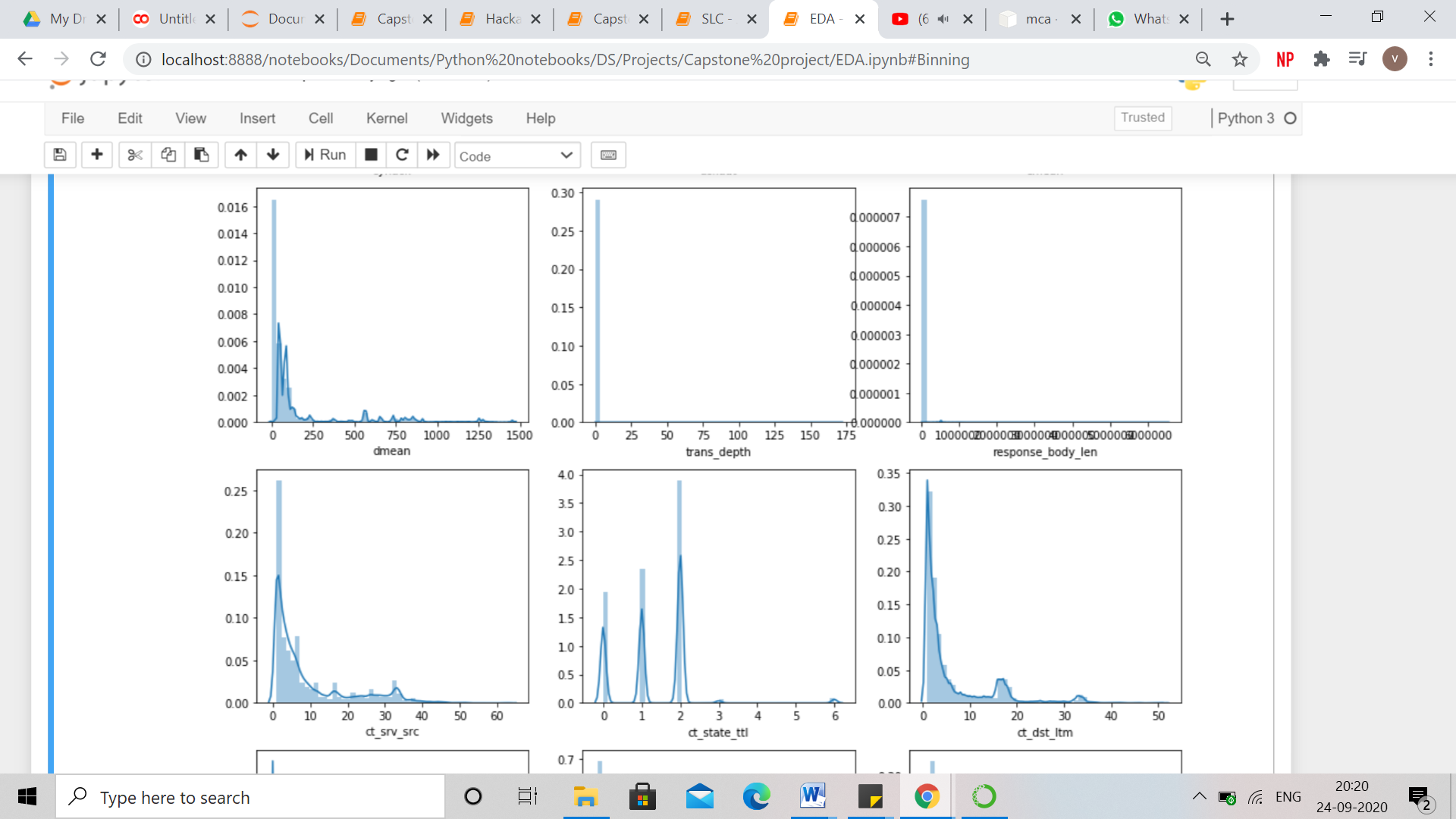
It can be observed that almost all the numerical features are right skewed and not normally distributed.

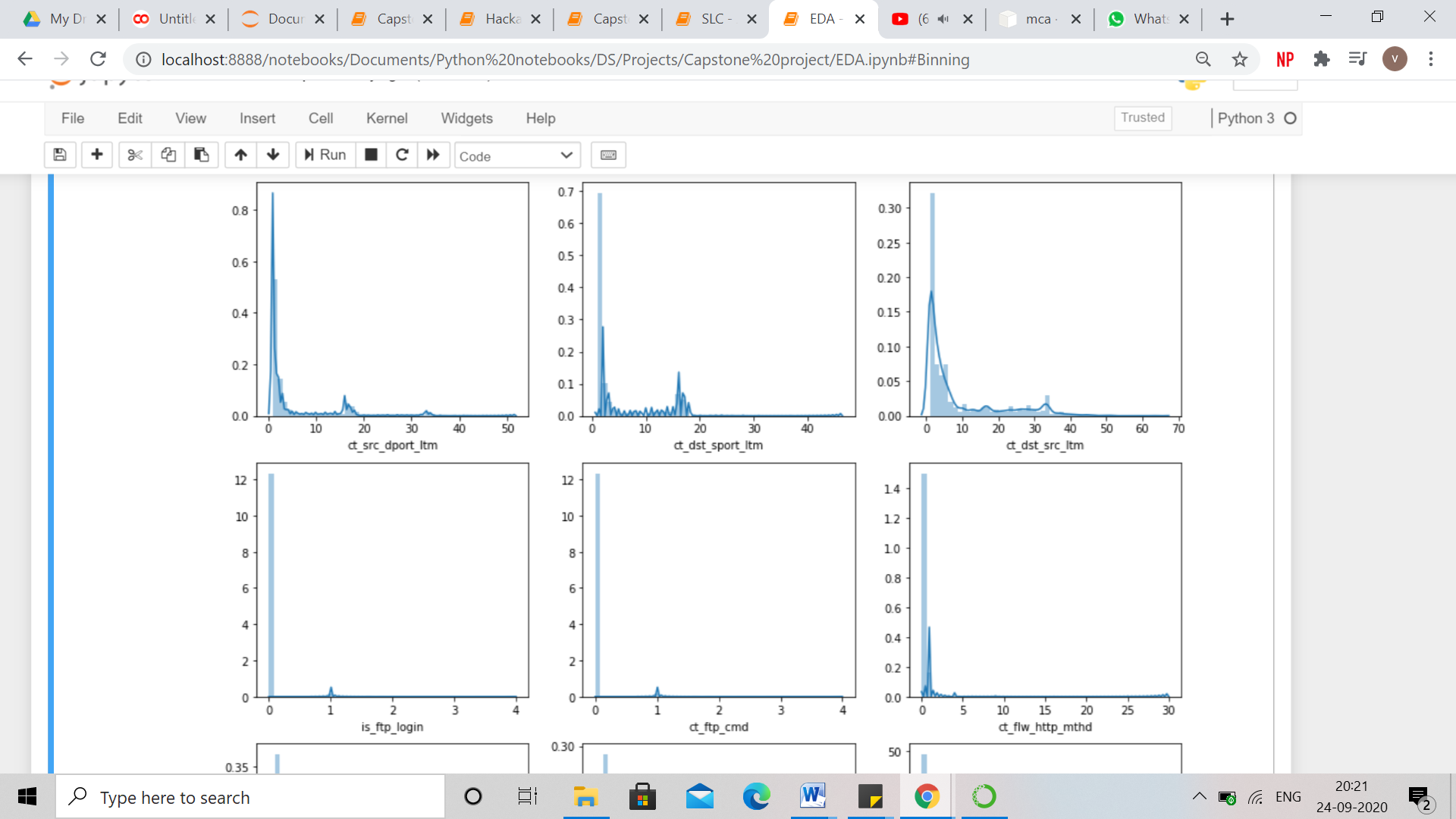


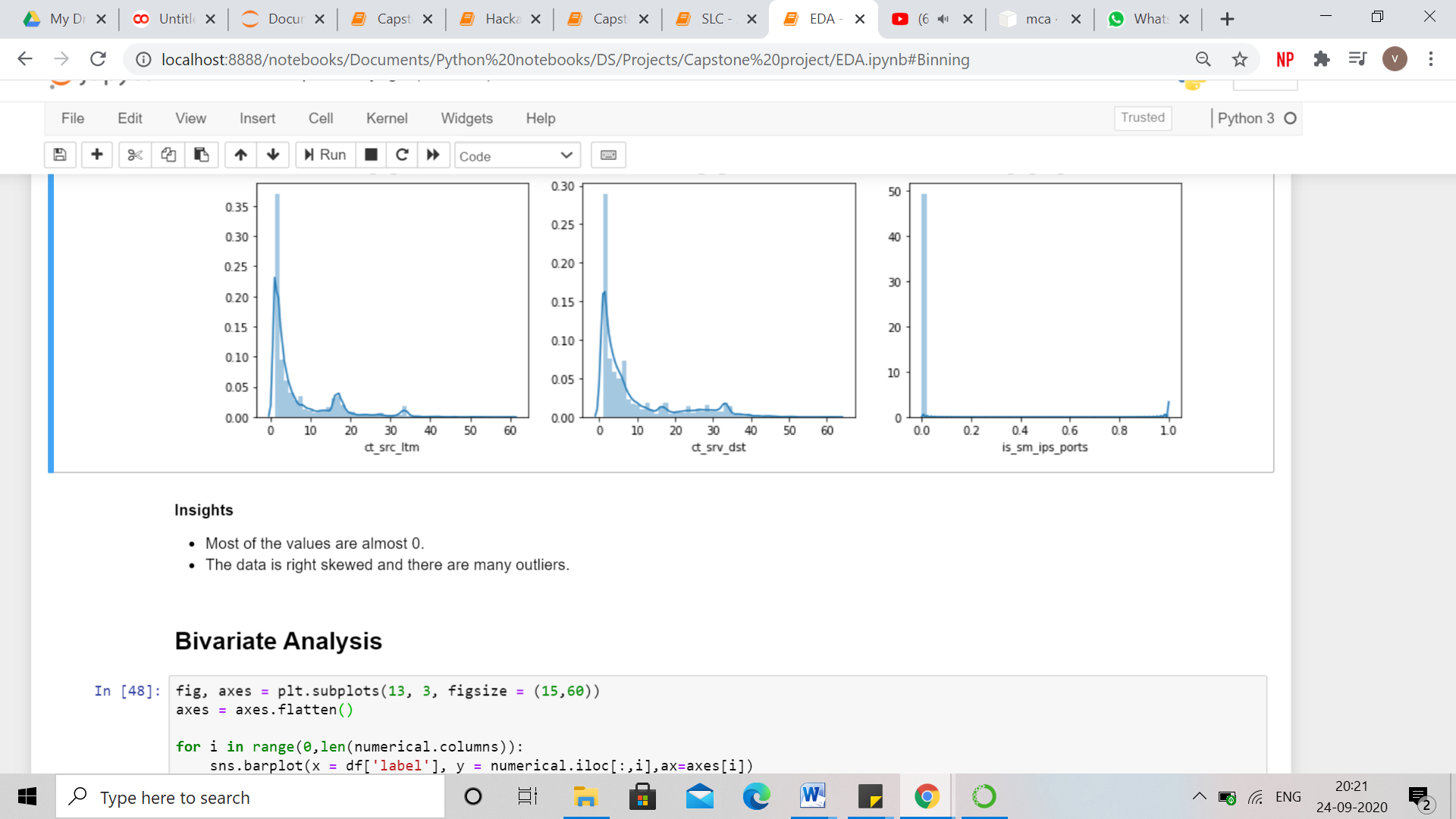












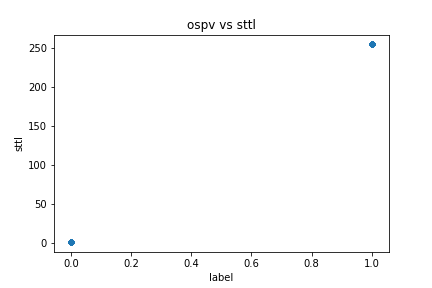
### 3.1.2 Presence of Outliers and its treatment

Outliers are stragglers — extremely high or extremely low values — in a data set that can throw off your stats. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

From the distribution plot of numerical features, it can be seen that the features are right skewed and hence there are outliers present at the right or higher side of the data. Higher values have lower frequency.

These outliers are a significant part of the data and help in understanding the patterns and making clusters and not a noise in the data. Therefore, these outliers are not treated and left as it is.

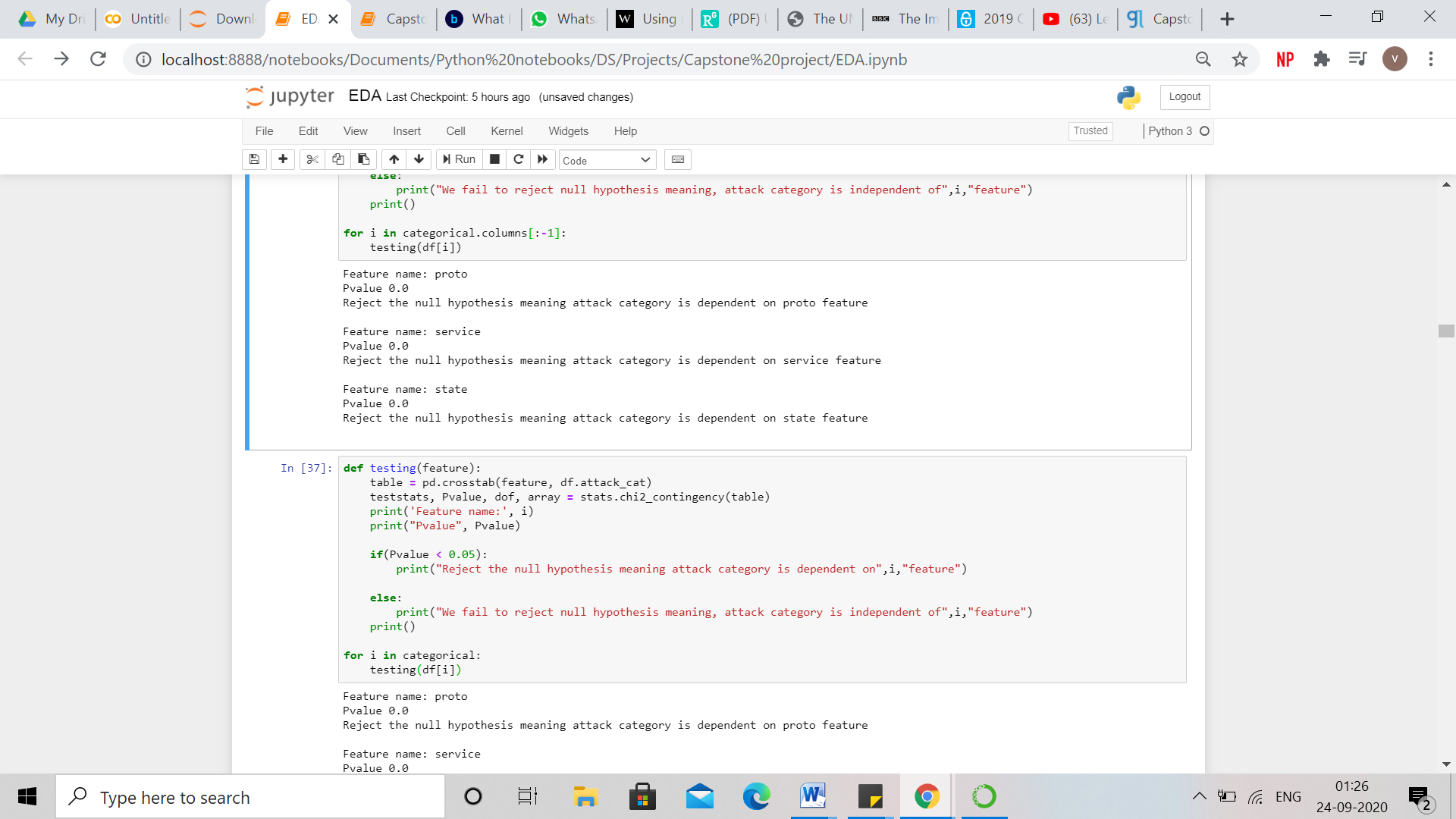
**Insights from bivariate analysis of both numerical and categorical features** –

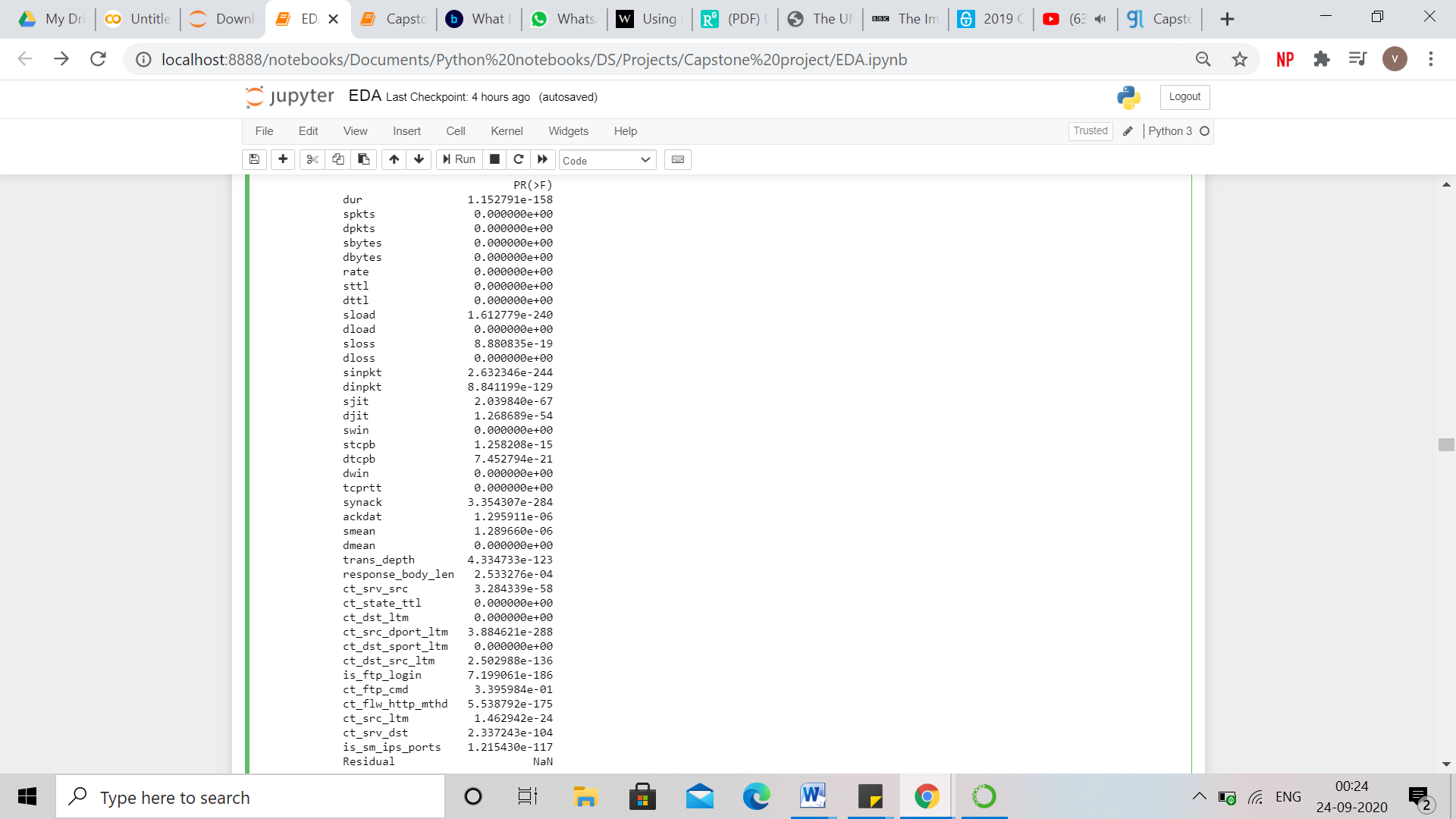
* When proto = 'unas', label = 1
* For trans depth of 3 and higher, all attacks are cyber-attacks.
* When is sm ips ports = 1, then there are no attacks.
* for state - ECO,URN,no all the attacks are 'Normal'
* for dns service, almost 80% of the attacks are Generic.
* for swin and dwin, all the values other than min value(0) and the max value(255), have 'Normal' attacks.
* For values higher than 25 in trans-depth, there are only cyber-attacks (label 1).
* for trans\_depth, 8 and 9 have 'DoS' attacks while 39,80,131,155,163 and 172 have 'Exploits' attacks
* for is\_sm\_ips\_ports, 1 has all the 'Normal' values while 0 has mixed attacks
* tcp is a more secure protocol than udp and unas because it has less attacks.
* udp proto has very high percentage of Generic attacks. It is highest among any other proto.
* UDP has high no. of Generic Attacks while TCP has high no. of Exploits and Fuzzers attacks.
* HTTP service using TCP proto has high number of attacks while DNS using UDP has high number of attacks.
* On an average, very high value of destination bits per second (dload) means there is no attack whereas a very low value of destination bits per second means that there is presence of a cyber-attack.
* In exploits and DoS cyber-attacks, on an average, number of packets sent from source to destination and vice versa is very high, the packets transferred is also very heavy and the data loss is also very high.
* On an average there are too many numbers of packets from source to destination (spkts) and vice versa (dpkts) for exploits and DoS attacks.
* Spread of duration (dur) is same across both the labels.
* When there is a cyber-attack, heavy data is transferred as both sbytes and dbytes have higher values as compared to non-attack values.
* Exploits and DoS attacks have very high values of sbytes. Same for dbytes. Heavy data is transferred in exploits and dos.
* Sloss values higher than 1000 represents exploits and DoS cyber-attacks.
* Data loss (sloss and dloss) is very high on an average in cyber-attacks.
* On an average there are very high values of destination bits per second for normal transactions. This means when there is no cyber-attack, data moves fast.
* Generic attack has lowest duration whereas Fuzzers has the highest duration. - Mean number of bits transferred per second is lowest for 'normal' and 'Worms' and highest for 'Generic'.
* On an average destination bits per second is less than 250,000 for all the attack categories except for 'normal'.
* Mean Source interpacket arrival time is less than 500 all attacks except for normal for which value is 2800.
* Mean Destiantion interpacket arrival time is highest for 'Fuzzer' whereas lowest for 'Generic'.
* Mean Source and Destination jitter time is lowest for 'Generic' , followed by 'Backdoor'.
* Mean TCP connection setup round-trip time is lowest for 'Generic' and highest for 'Worms'.
* High values of synack and ackdat represents normal attacks.
* When there are 'Worm' attacks and service= not used then proto = udp and when there are 'Worm' attacks and service= http then proto = tcp.
* When we use 'ospf' protocol with very less time to live in network, most likely the transactions being normal whereas when proto = ospf and sttl is maximum then there is attack.
* 
* There is no data loss in ospf protocol
* Whenever dns service is used with TCP protocol, no transaction is normal.
* When both is\_ftp\_login and ct\_ftp\_cmd is 1 for all 'Normal','Fuzzers','Exploits','DoS' attack
* While using 'ospf' protocol, no 'Worms' attacks happened.
* There is no analysis attack when using udp protocol.
* There is very less likelihood(less than 2%) that dos attack will occur in tcp and udp protocol.
* When using udp, around 60% of the times, generic occurs.
* Shellcode and worm occurs only in tcp and udp .
* there is more chance for no attack when state is con compared to other states
* Generic mosty occurs in int state.
* when dns is the service, around 85% of the times attack will be generic
* Shellcode is there only when service = not\_used that means for dns, http and other service,there are no shellcode attacks.
* Worms occurs only in not used and http , ratio being 1:9
* When the service is dns, then there is a very high possibility that the attack will be Generic.

### 3.1.3 Statistical Significance of variables

Statistical significance is the likelihood that a relationship between two or more variables is caused by something other than chance.

Statistical significance is used to provide evidence concerning the plausibility of the null hypothesis, which hypothesizes that there is nothing more than random chance at work in the data. Statistical hypothesis testing is used to determine whether the result of a data set is statistically significant.





Statistical Significance of numerical features has been tested by ANOVA and for categorical features, chi-square has been used.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

The null hypothesis states that the feature does not have any significant effect on the result/target feature. Whereas, the alternate hypothesis states that at least one of the sample means is different from another and the feature has a significant effect on the target feature.

The Chi Square statistic is commonly used for testing relationships between categorical features. The null hypothesis of the Chi-Square test is that no relationship exists between the categorical features i.e. they are independent of each other and whereas alternate hypothesis states that there is a relationship between the categorical features and independent feature has an effect on the target variable.

In both the cases, p-value is less than the alpha (default = 0.05) which states that null hypothesis has been rejected and the features are statistically significant to predict the label - target variable.

Significance testing was also done for all the features with respect to each attack category. Below are the insights inferred using the p-value rule -

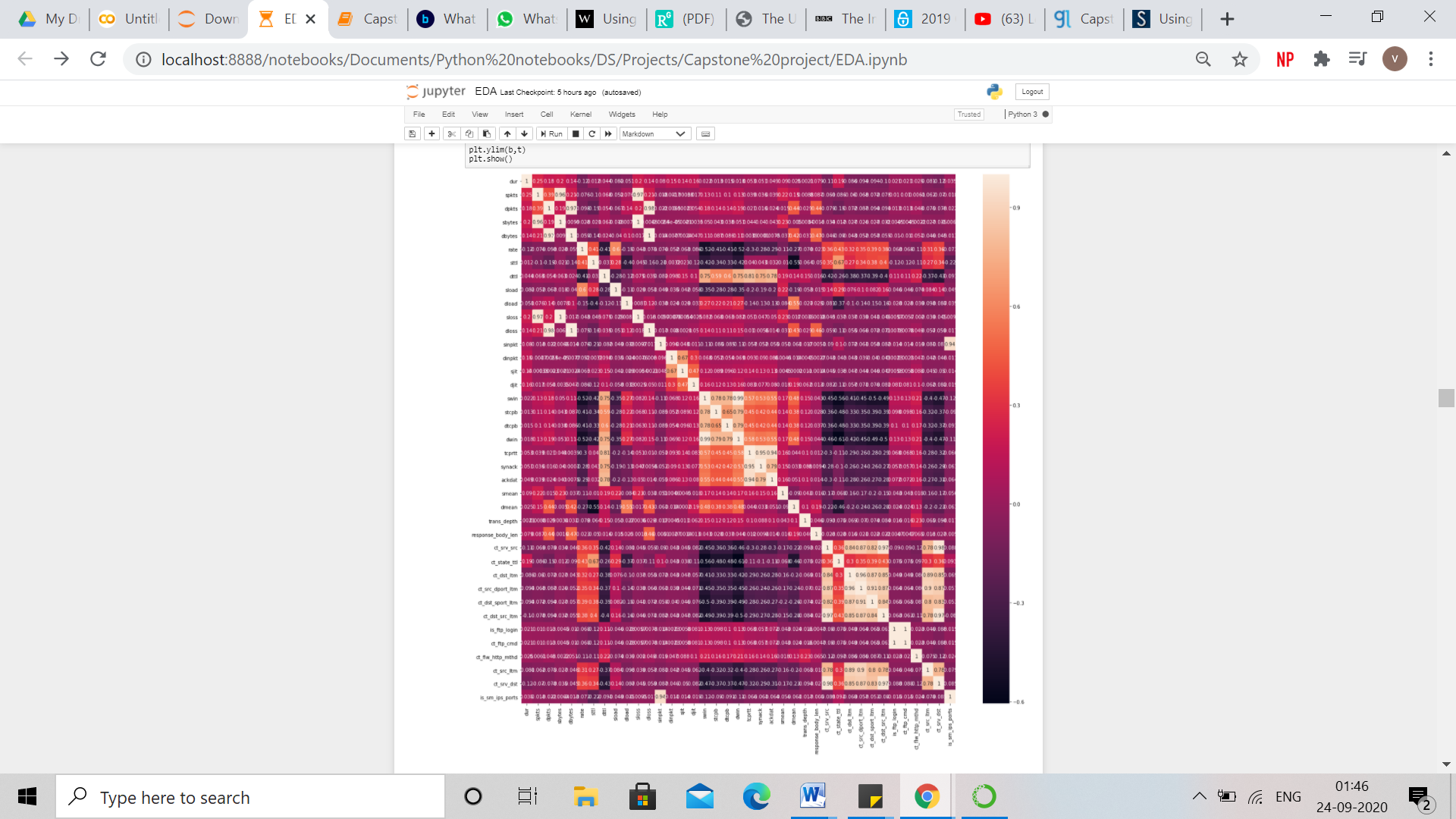
* ackdat, dloss and response body length features have no role in determining the attack category "Analysis".
* Sjit, synack, ackdat, sbytes, response body length and dloss features have no role in determining the attack category "Backdoor".
* sjit, ackdat, spkts,smean and response body length features have no role in determining the attack category "DoS".
* tcprtt, dpkts and dbytes features have no role in determining the attack category "Exploits".
* tcprtt, dpkts and dbytes features have no role in determining the attack category "Fuzzers".
* ackdat,dloss and response body length features have no role in determining the attack category "Generic".
* Dinpkt, ackdat,sjit features have no role in determining the attack category "Reconnaissance".
* Sjit, rate, ackdat, dinpkt, sbytes, sloss and response body length features have no role in determining the attack category "Shellcode".
* Sjit, dur, dinpkt, ackdat, spkts, dbytes, sloss dtcpb , smean and djit features have no role in determining the attack category "Worms".

### 3.1.4 Multicollinearity

Multicollinearity occurs when [independent variables](https://statisticsbyjim.com/glossary/predictor-variables/) in a dataset are correlated. This [correlation](https://statisticsbyjim.com/glossary/correlation/) is a problem because independent variables should be independent so that there would be no noise acting in the data. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

The problem of multicollinearity is different in clustering analysis because there’s no dependent variable or beta coefficient. A certain number of observations measured on a specified number of variables are used for creating segments. Each observation belongs to one segment, and each segment can be defined in terms of all the variables used in the analysis. When variables used in clustering are collinear, some variables get a higher weight than others. If two variables are perfectly correlated, they effectively represent the same concept.

One major use of PCA lies in overcoming the multicollinearity problem. PCA can aptly deal with such situations by excluding some of the low-variance principal components.



# **4. FEATURE ENGINEERING**

# **4.1 Whether any transformations required**

The UNSW-NB dataset is a network sever data. Unlike the usual datasets, this data does not require outlier treatment due to which there is no need of transformation also. Hence, we have not done any kind of transformation on the data.

# **4.2** Scaling the data

We have scaled the numerical data using the StandardScaler available in Sklearn library.

Below are few points explaining StandardScaler:

**Where does it come from?**

It comes from [scikit-learn](http://scikit-learn.org/stable/), which is a free machine learning library for Python. Working with data science problems you will use scikit-learn a lot. It’s a simple and efficient tools for data mining and data analysis.

**What does it do?**

It will resize your features. It assumes that the data for each feature is normally distributed and will scale the feature so that the distribution is now centered around 0 and has a standard deviation of 1.

**When can it be used?**

If the range of values in the features varies widely.

**Why is it important?**

Especially for distance based models scaling is important. If one of the features has a broader range of values than the other features, the overall distance will be heavily affected by this feature. Therefore, the range of all features should be scaled so that each feature contributes approximately proportionately to the final distance.

**4.3 Feature selection**

We have done feature selection on basis of centroids of features which are distinct and significant in cluster making process. Below are few of the features which we dropped along with their reasons:

* Id: - Id column contains all unique values/ serial number of each row which is of no use in our analysis. Hence we have dropped Id column.
* Tcprtt: It is basically the sum of 2 columns i.e. synack and ackdat. In order to remove redundancy from the data, we have removed this column.
* ct\_ftp\_cmd : Majority of data of this column is similar to the data of is\_ftp\_login column. Hence, in order to avoid multicollinearity, we have removed this column. Below is the snapshot of the same.

**4.4 Dimensionality reduction**

We have used PCA (Principal Component Analysis) for dimensionality reduction on our numerical features. Below is the detailed explanation of this feature extraction technique i.e PCA:

PCA is a very useful technique that can help de-noise and detect patterns in data. As a summary, what PCA do is to combine every feature and to extract the top more relevant ones automatically. It is a systemized way to transform input features into principal components, and uses them as the new features.

Principal components are directions in data that maximize variance (minimize information loss) when projecting or compressing them down.

The more the variance of data along a Principal Component, the more information that direction contains and the higher that principal component is ranked. The number of Principal Components will be less or equal the number of input features.

Sci-kit learn library offers a powerful PCA component classifier. This code snippet illustrates how to create PCA components:

# Understanding PCA

This section of the article provides an overview of the process:

* PCA technique analyses the entire data set and then finds the points with maximum variance.
* It creates new variables such that there is a linear relationship between the new and original variables such that the variance is maximized.
* Covariance matrix is then created for the features to understand their multi-collinearity.
* Once the variance-covariance matrix is computed, PCA then uses the gathered information to reduce the dimensions. It computes orthogonal axes from the original feature axes. These are the axes of directions with maximum variance.

Firstly the eigenvectors of the variance-covariance matrix are calculated. The vector represents the directions of maximum variance which are known as the principal components. The eigenvalues are then created that define magnitude of the principal components.

The eigenvalues are the PCA components.

Therefore, for N dimensions, there will be a NxN variance-covariance matrix and as a result, we will have a eigen vector of N values and N eigen values matrix.

We need to take the eigen vectors that represent the our data set best. These are the vectors which we have highest eigenvalues.

NOTE: If we want to keep sci-kit learn to give us all of the PCA components so that we can assess the variance then initialize PCA with None components. Also, it is important to normalize/standardize the data before performing PCA because PCA is sensitive to the scale of the data in the features.

# **5. Model Building**

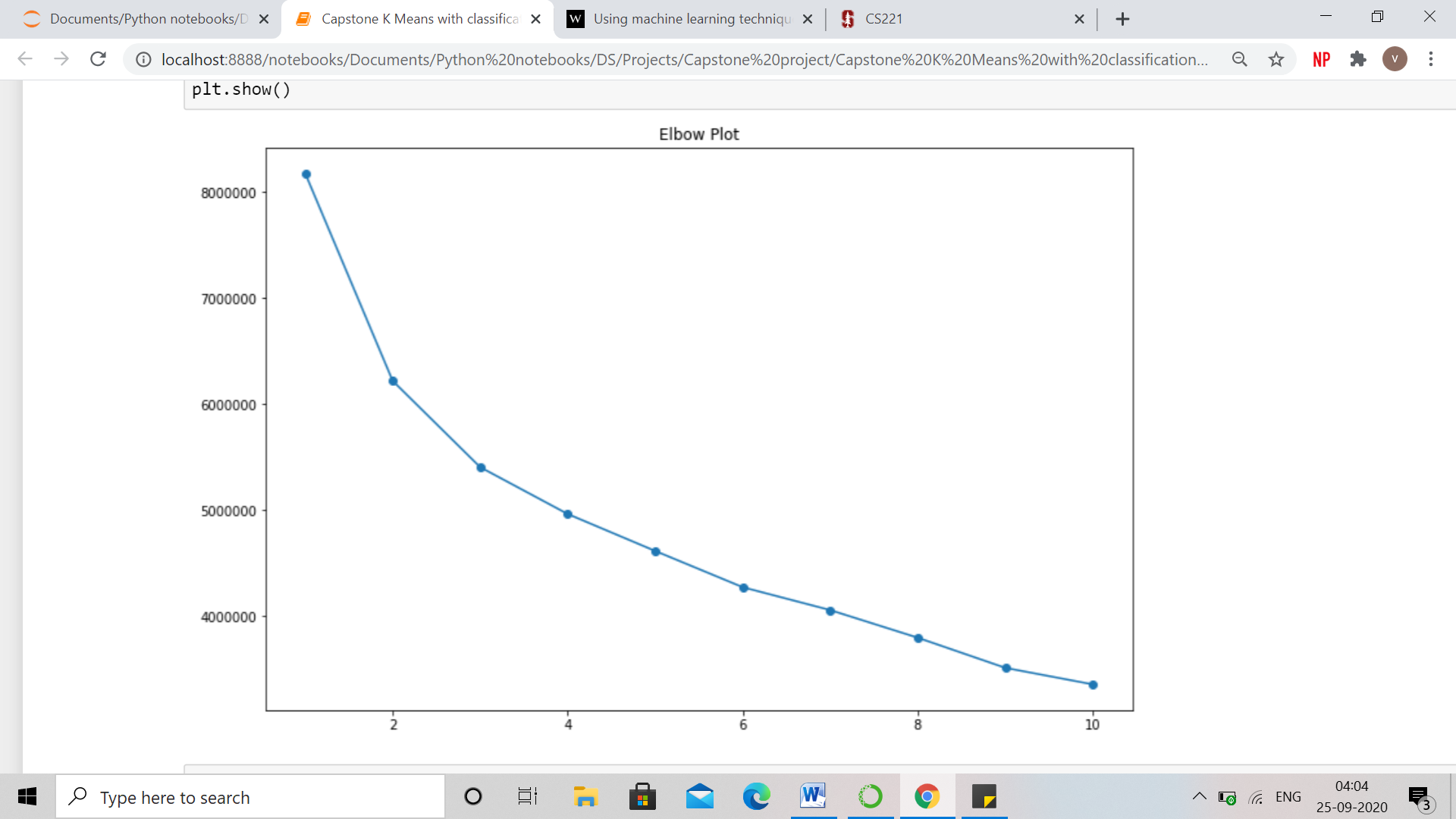
# **5.1 Clustering**

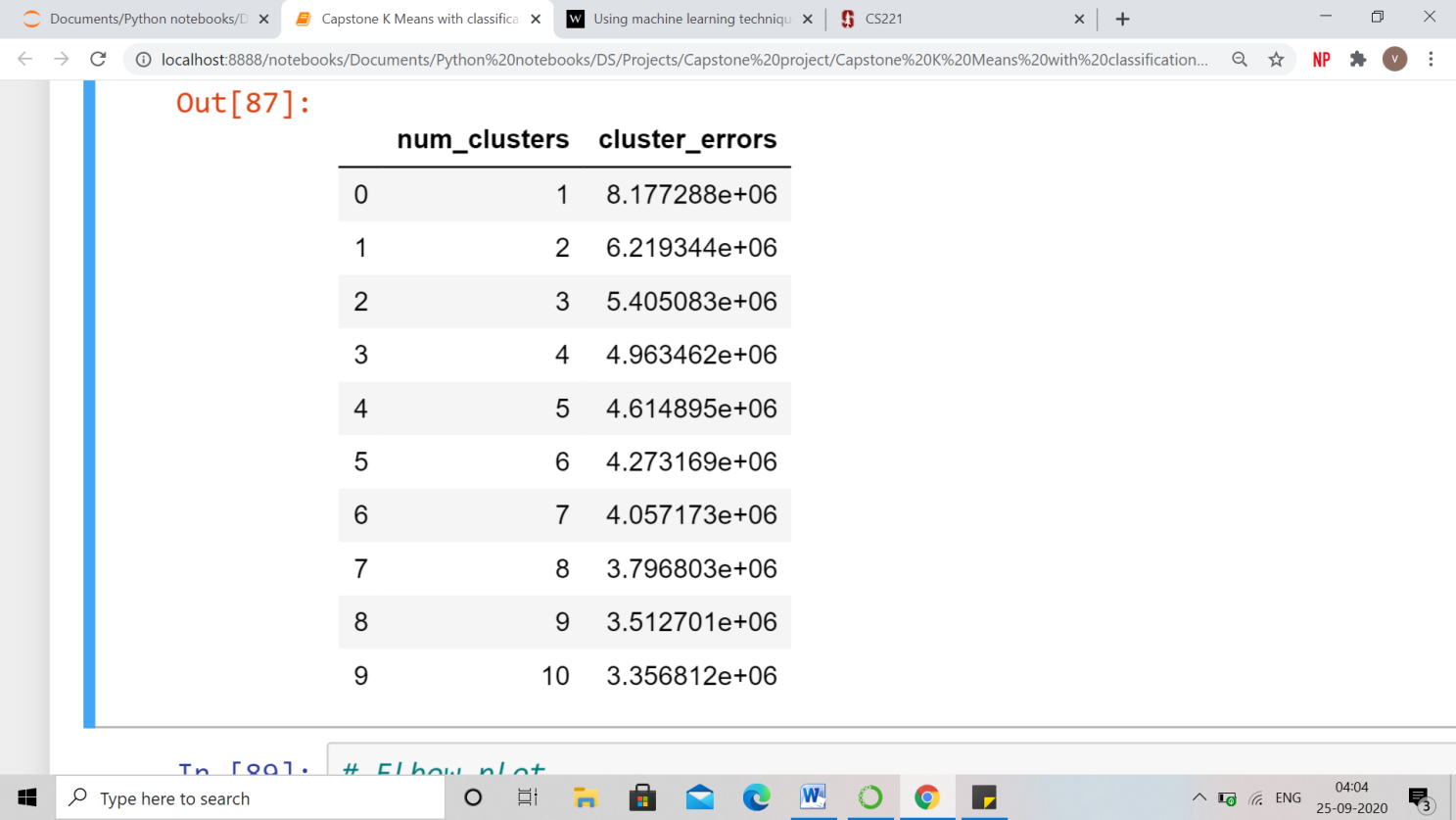
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

For clustering K-Means clustering algorithm is being used for this dataset. K-Means is one of the most popular "clustering" algorithms. K-means stores ‘k’ centroids that it uses to define clusters. A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.

For k – means clustering algorithm:

* First of all, find the best k value
* Then use this k value to create a k-means model.
* Choosing the best value for k:

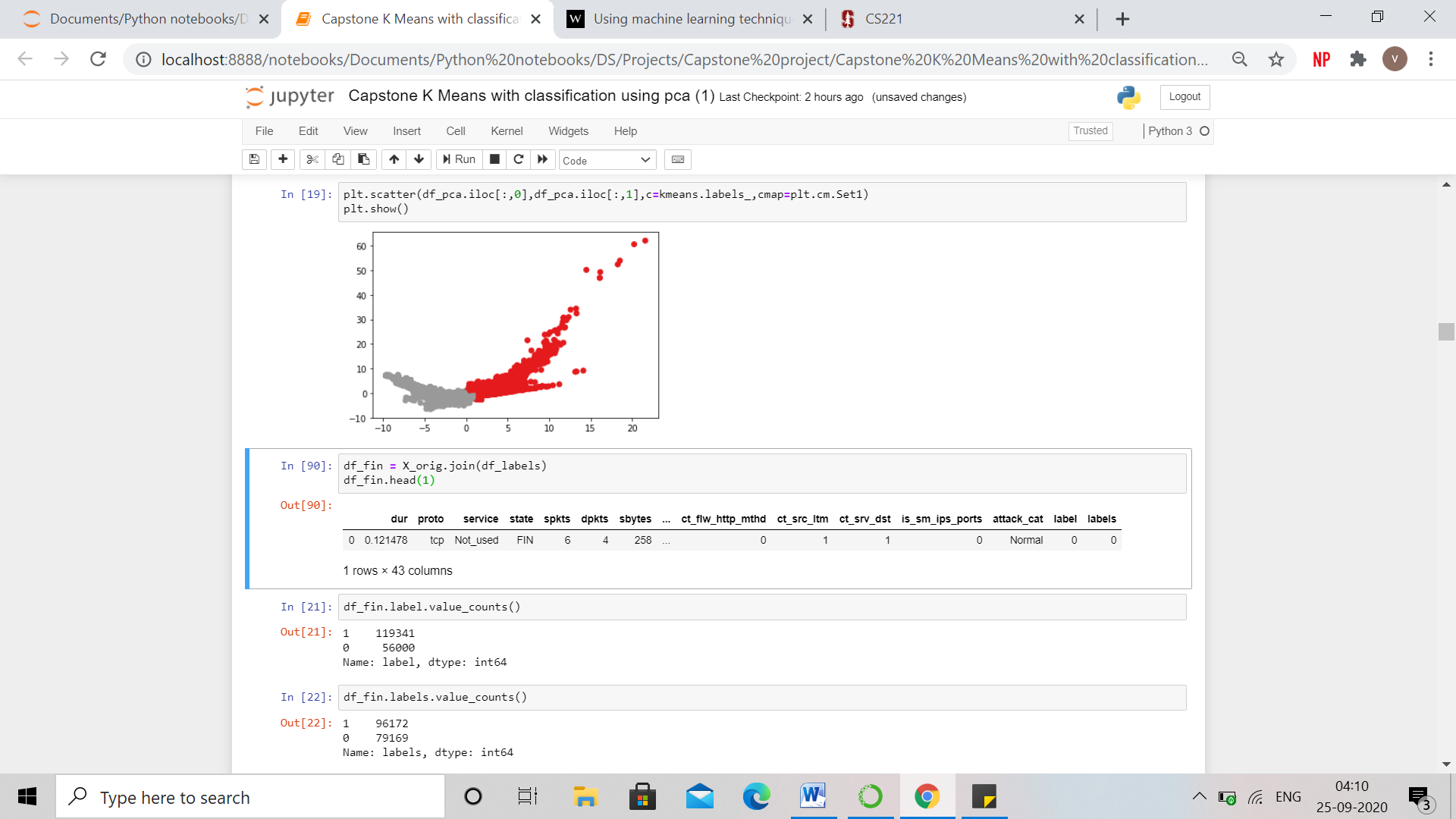


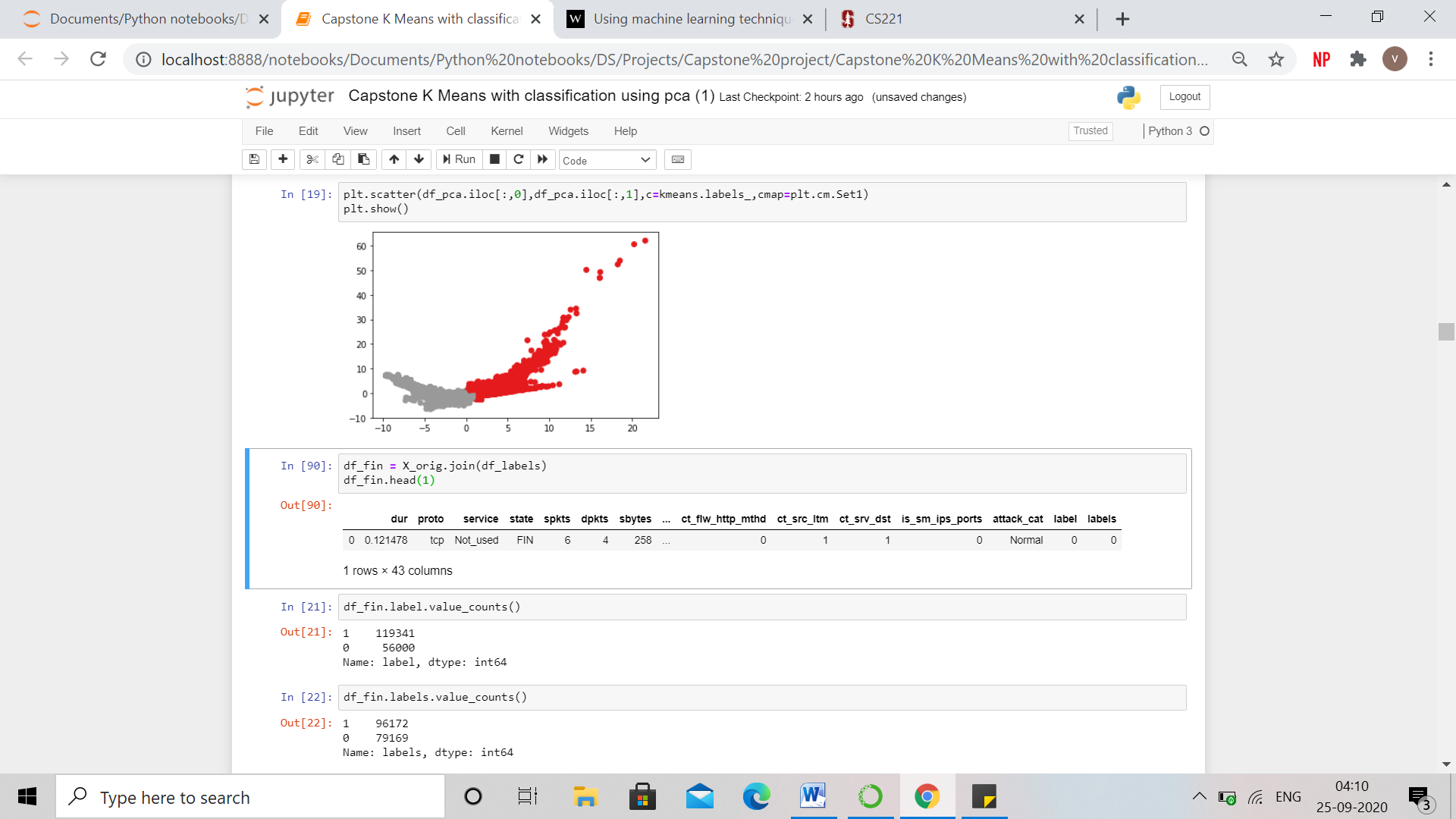


By looking at elbow plot and inertia values, there is a greater fall at 2.

### 5.1.1 k = 2

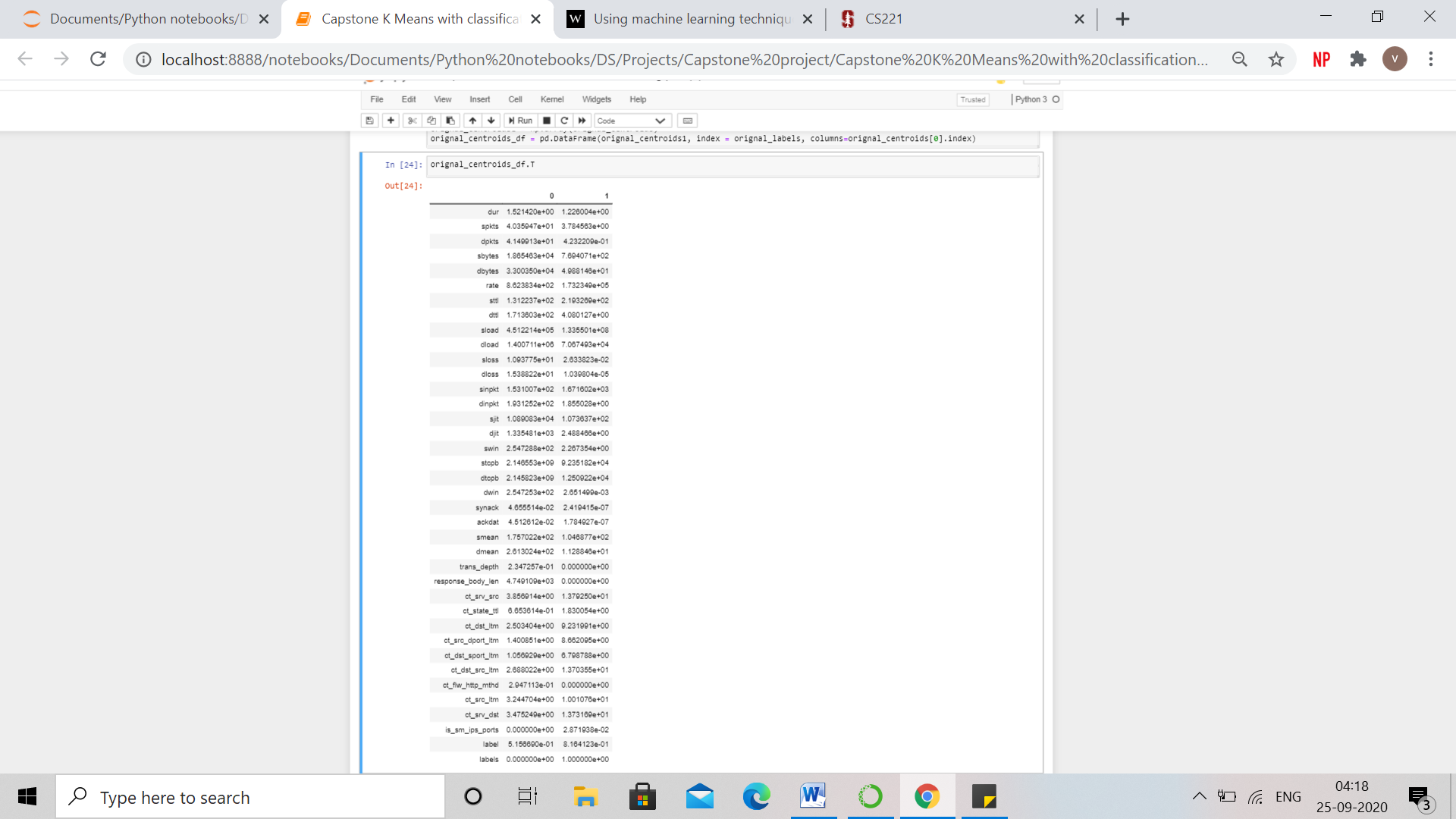
Below is the scatter plot representing 2 clusters and the value counts of the labels obtained from clustering. On the dataset pca has been applied.





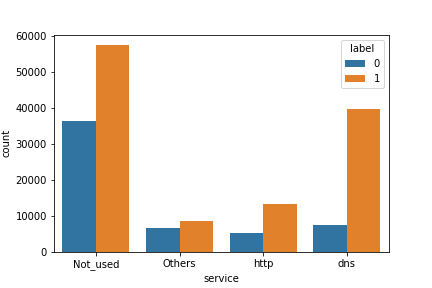
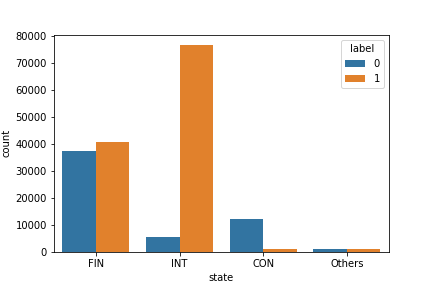
Below is the centroid values obtained from the clustering. For few of the features, centroids values are very different for different labels.

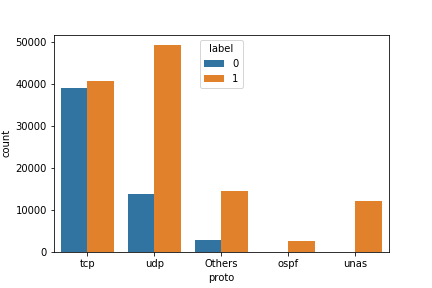
These features are - dpkts, sbytes, dbytes, rate, sloss, stcpb, dtcpb, is\_sm\_ips\_ports, ct\_srv\_src



From the clustering few patterns can be observed –

|  |  |  |
| --- | --- | --- |
|  | Non Attack (Cluster 0) | Attack (Cluster 1) |
| proto | tcp | udp |
| state | FIN | INT |
| dpkts | 41.5 | 0.423 |
| sbytes | 18654.63 | 769.407 |
| dbytes | 33003.5 | 49.88 |
| rate | 862.38 | 173234.9 |
| sloss | 10.93 | 0.02633 |
| stcpb | 2.146553e+09 | 9.235182e+04 |
| dtcpb | 2.145823e+09 | 1.250922e+04 |
| is\_sm\_ips\_ports | 0 | 0.0287 |
| ct\_srv\_src | 3.857 | 13.8 |

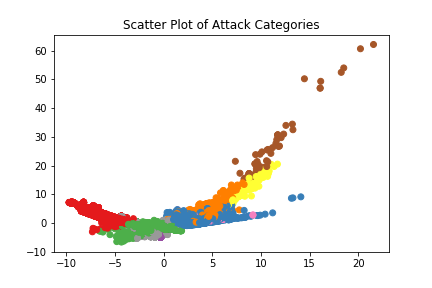
 



### 5.1.2 k =10

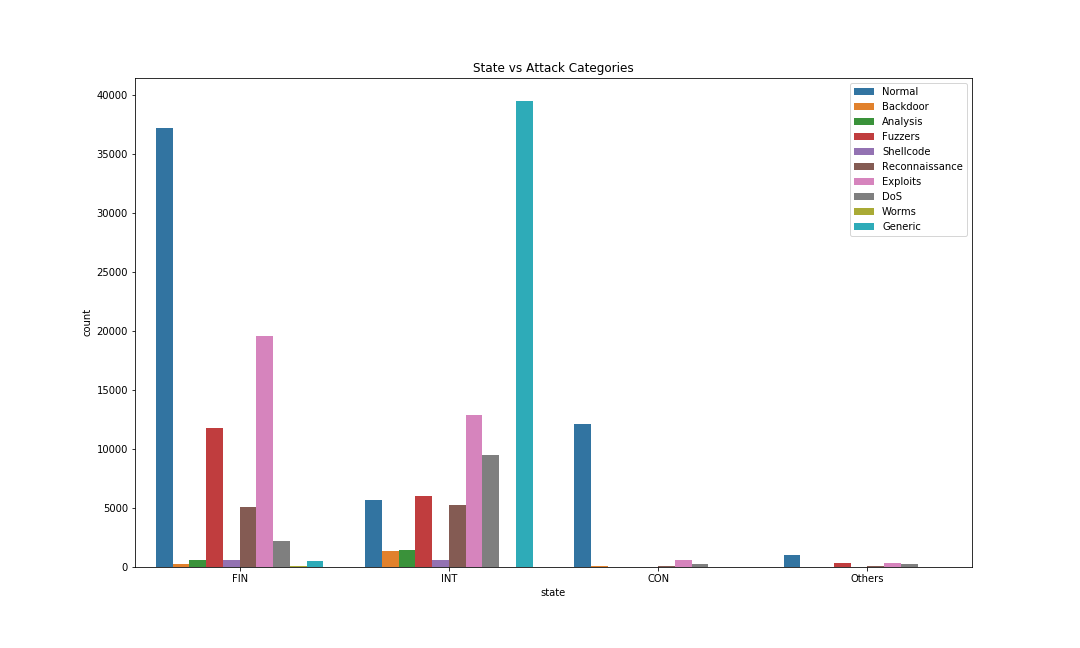
Now, value of k is chosen as 10, because as per prior knowledge there are 10 attack categories.

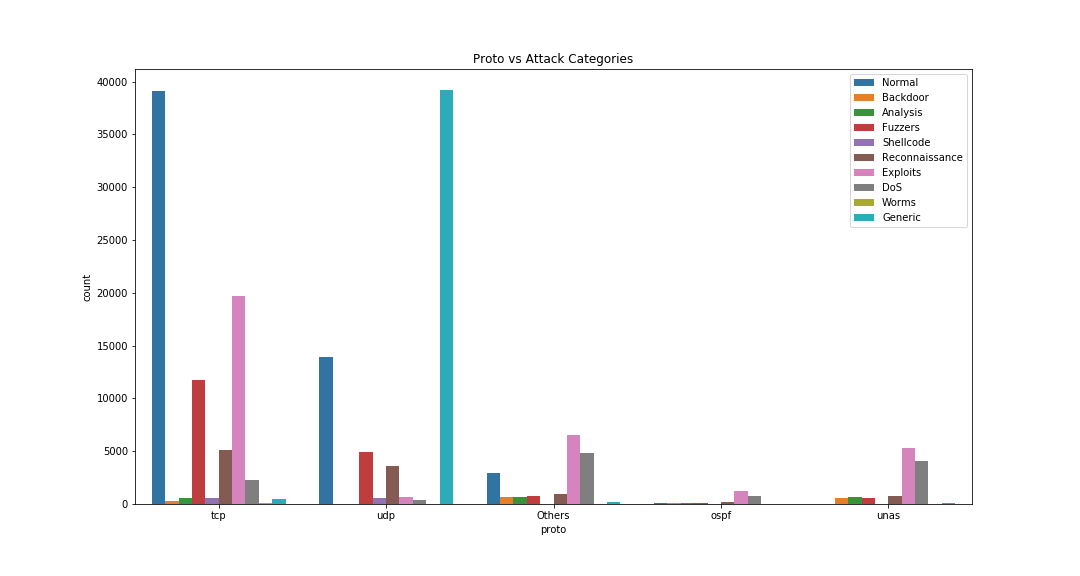
Below, is the scatter plot for the same –

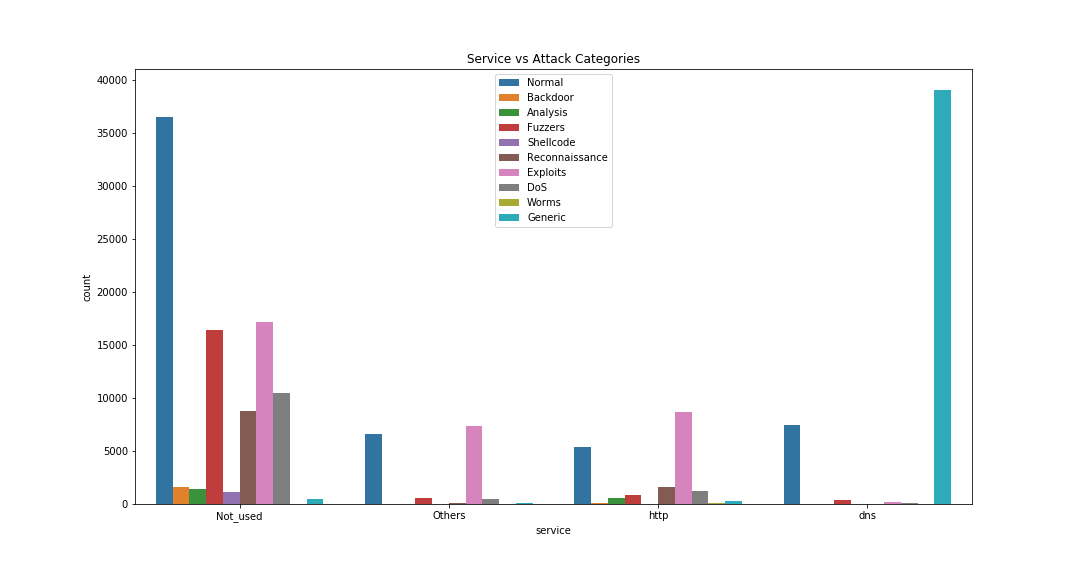


As it can be seen from the scatter plot, the clusters are not very much separable and clear.

Some of the features that were still able to separate the clusters are – dur, ct\_state\_ttl , sbytes , dpkts and rate.







# **5.2 Classification**

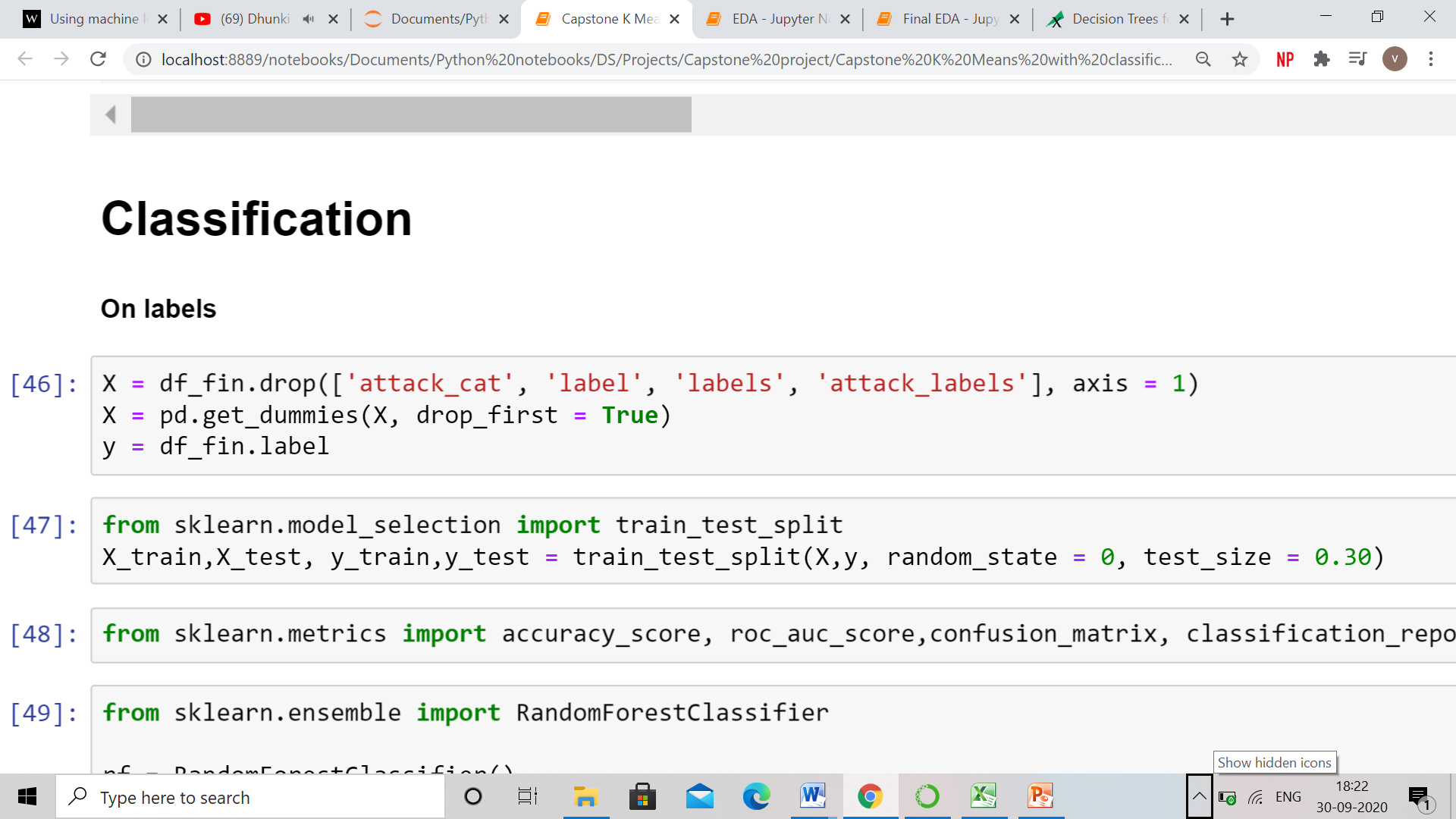
A classification model was built for 2 and 10 clusters and the following algorithms were used -

**Naive Bayes** classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

**Decision Trees** are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

**Random Forest Classifier** is an ensemble tree-based learning algorithm. It is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.

The dataset was split into train and test in the 70:30 ratio.



Several performance measures, i.e. accuracy, precision, recall and false alarm rate as calculated as follows.

* Accuracy = (TP+TN)/ (TP+FP+TN+FN)
* Precision =TP /(TP+FP)
* Recall =TP /(TP+FN )
* Sensitivity or True Positive Rate (TPR)=TP /(TP + FN)
* Specificity or True Negative Rate (TNR)=TN /(FP + TN)
* FPR =FP /(FP+TN)
* FAR =(FPR + FNR) /2
* FNR = FN /(FN+TP)
* F1 Score = 2(Precision x Recall)/(Precision +Recall)

where

* True positive(TP) means the correct intrusion detection
* False Positive(FP) means to assume the normal traffic as the cyberattack.
* True negative (TN) refers to normal traffic correctly labeled as normal.
* False Negative (FN) means to fail intrusion disclosure.
* FPR is the false positive rate.
* FNR is the false negative rate.
* False Alarm Rate (FAR) means the average ratio of the misclassified to classified records either normal or abnormal.
* The F1score refers the harmonic average of the precision and recall.

### 5.2.1 Classification model for Label Category

Both of the algorithms are applied on 3 datasets – original dataset without pca, dataset with pca and then features selected from clustering.

**From Random Forest –**

* Original dataset without PCA : Accuracy Score = 0.96
* Original Dataset with PCA : Accuracy Score = 0.95
* Selected Features - proto, state, service, dpkts, sbytes, dbytes, rate, sloss, stcpb, dtcpb, is\_sm\_ips\_ports, is\_ftp\_login, ct\_state\_ttl, dur, ct\_srv\_src

Accuracy Score = 0.9533

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Original dataset without PCA** | | **Original dataset with PCA** | | **Selected Features without PCA** | |
|  | Precision | Recall | Precision | Recall | Precision | Recall |
| Attack | 0.96 | 0.98 | 0.95 | 0.98 | 0.95 | 0.98 |
| NoAttack | 0.96 | 0.92 | 0.96 | 0.98 | 0.95 | 0.90 |

As, we are getting almost same precision, recall and accuracy even after removing so many features, that means the dataset can be made compact. It also shows that these features are enough to determine whether it is an attack or not.

**From Naïve Bayes –**

* Original Dataset without PCA

Accuracy Score = 0.8676

|  |  |  |
| --- | --- | --- |
|  | **Precision** | **Recall** |
| **Attack** | 0.84 | 0.86 |
| **NoAttack** | 0.69 | 0.66 |

**From Decision Tree –**

* Original Dataset without PCA

Accuracy Score = 0.9473

|  |  |  |
| --- | --- | --- |
|  | **Precision** | **Recall** |
| **Attack** | 0.96 | 0.96 |
| **NoAttack** | 0.92 | 0.92 |

From the above 3 models, Random Forest has the highest, accuracy, precision and recall score. Hence, for classification of different attack categories, we will be using Random Forest Classifier only.

**Performance Measures for Random Forest Classifier of Selected Features –**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Predicted Class** | |
|  |  | **Positive** | **Negative** |
| **Actual Class** | **Positive** | 35157 | 754 |
| **Negative** | 1707 | 14985 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Value** | **Metric** | **Value** |
| TP | 35157 | Sensitivity (TPR) | 0.979 |
| FP | 1707 | Specificity (TNR) | 0.8977 |
| TN | 14985 | FPR | 0.1022 |
| FN | 754 | FNR | 0.021 |
| FAR | 0.061 | F1 Score | 0.95 |

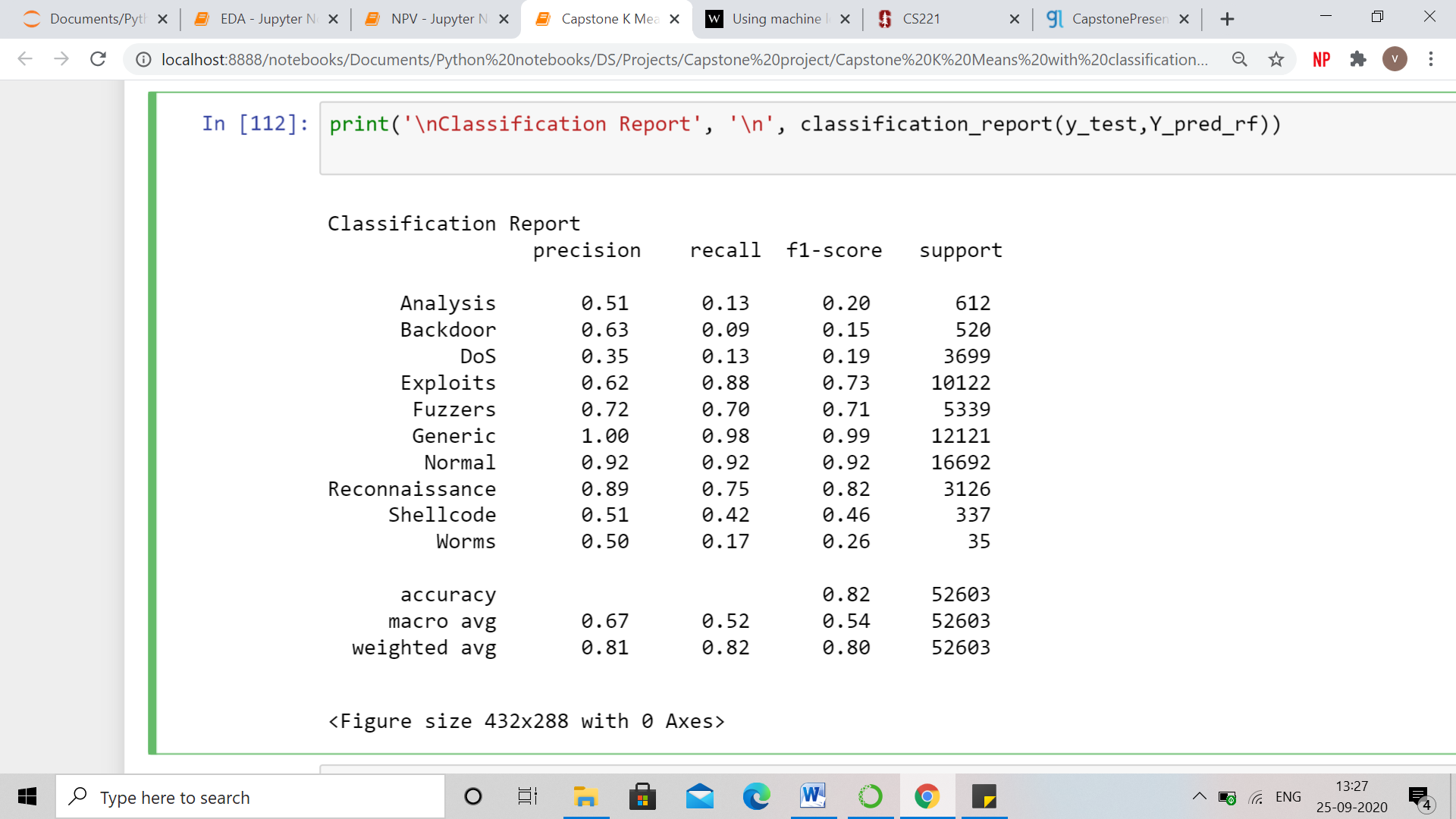
For our problem, FNR should be lower than FPR

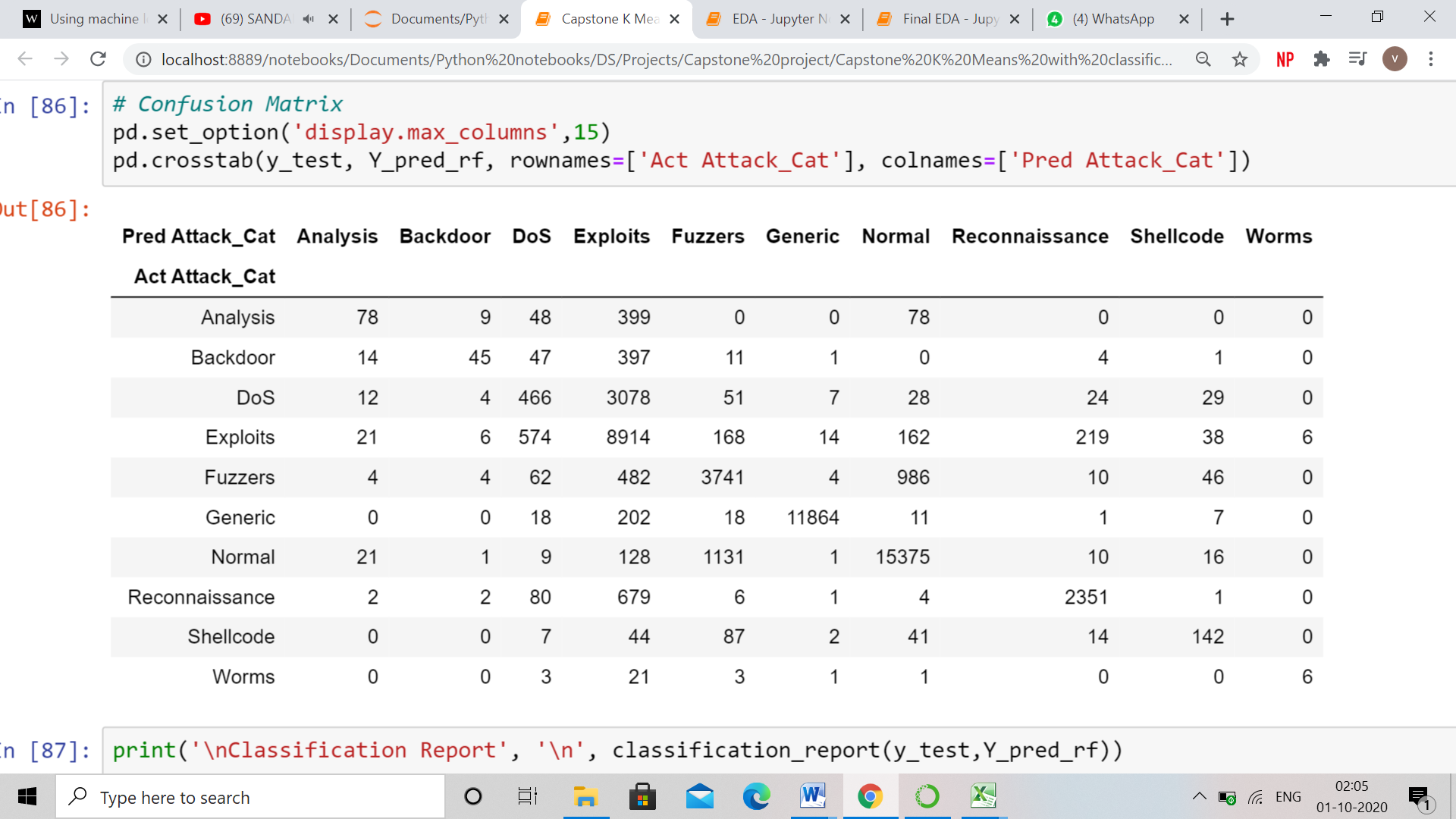
### 5.2.2 Classification Model for different attacks

**Random Forest Classifier was applied to classify the attacks**.

* Original dataset without PCA : Accuracy Score = 0.8206
* Original Dataset with PCA : Accuracy Score = 0.79
* Selected Features - proto, state, service, dpkts, sbytes, dbytes, rate, sloss, stcpb, dtcpb, is\_sm\_ips\_ports, is\_ftp\_login, ct\_state\_ttl, dur, ct\_srv\_src

Accuracy Score = 0.8171



**Confusion Matrix**

This is the classification report derived by using the selected features. Random forest was able to classify ‘Exploits’, ‘Fuzzers’, ‘Generic’, ‘Normal’ and ‘Reconnaissance’ well based on precision and recall score.

‘Analysis’, ‘Backdoor’, ‘DoS’ and ‘Worms’ have low precision or recall score. One of the reasons of this could be imbalance distribution of attack categories.

**Performance Metrics of Different Attack Categories**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **TP** | **FP** | **TN** | **FN** | **TNR** | **TPR** | **FNR** | **FPR** | **FAR** |
| **Analysis** | 78 | 74 | 42904 | 534 | 0.9983 | 0.1275 | 0.8725 | 0.0017 | 0.4371 |
| **Backdoor** | 45 | 26 | 42937 | 475 | 0.9994 | 0.0865 | 0.9135 | 0.0006 | 0.4570 |
| **DoS** | 466 | 848 | 42516 | 3233 | 0.9804 | 0.1260 | 0.8740 | 0.0196 | 0.4468 |
| **Exploits** | 8914 | 5430 | 34068 | 1198 | 0.8625 | 0.8815 | 0.1185 | 0.1375 | 0.1280 |
| **Fuzzers** | 3741 | 1475 | 39241 | 1598 | 0.9638 | 0.7007 | 0.2993 | 0.0362 | 0.1678 |
| **Generic** | 11864 | 31 | 31118 | 239 | 0.9990 | 0.9803 | 0.0197 | 0.0010 | 0.0104 |
| **Normal** | 15375 | 1311 | 27607 | 1317 | 0.9547 | 0.9211 | 0.0789 | 0.0453 | 0.0621 |
| **Reconnaissance** | 2351 | 282 | 40631 | 774 | 0.9931 | 0.7523 | 0.2477 | 0.0069 | 0.1273 |
| **Shellcode** | 142 | 138 | 42840 | 195 | 0.9968 | 0.4214 | 0.5786 | 0.0032 | 0.2909 |
| **Worms** | 6 | 6 | 42976 | 29 | 0.9999 | 0.1714 | 0.8286 | 0.0001 | 0.4144 |

# **6. Conclusion**

In an attempt to find similarity or dissimilarity among the attack categories, attack and no attack were easily separable and easily classified but there was less distinction between the different attack categories. Few of the conclusions that can be drawn are -

* While using TCP protocol along with 'dns' service, there are more chances of attack being happened.
* While using 'ospf' protocol with less time to live in network, most likely the transactions being normal.
* Most of the attack happens while using 'udp' protocol
* While using 'ospf' protocol, no 'Worms' attacks happened but no transaction was normal. There were no worm attacks where protocol is ospf because Worms are malware that replicate themselves and spread to other computers by using the network to spread the attack but as ospf follows the most efficient shortest path, it is very difficult to replicate the malware in network.
* When is\_sm\_ips\_ports = 1, then there are no attacks.
* TCP is a more secure protocol than udp and unas .
* From clustering of attack / no attack, it can be conclude that attack mostly occurs when udp is the protocol, service is either no used or dns, INT is the state and high values of dpkts, sbytes, dbytes, sloss, stcpb, dtcpb and low value of dload, djit, synack and ackdat.

Cyber-attacks can be distinguished from each other on the following basis –

* **Exploits** - In exploits cyber-attacks, on an average, number of packets sent from source to destination and vice versa is very high, the packets transferred is also very heavy and the data loss is also very high. And High values of trans depth represents exploits attack
* **Analysis** - There is no analysis attack when using udp protocol.
* **DoS** - In exploits cyber-attacks, on an average, number of packets sent from source to destination and vice versa is very high, the packets transferred is also very heavy and the data loss is also very high.
* **Generic** – When service is dns, protocol is udp and state is int, then 99% attacks are Generic. On an average, interpacket arrival time (dinpkt and sinpkt) is least for Generic attack.
* **Shellcode** attack can only occur when there is no service and protocol is either tcp or udp.
* **Worms** - No worm attacks happen when using ‘ospf’ protocol and they occur when there is no service.

Exploits, Generic and Normal were interpretable from the clusters. We were able to better classify ‘Exploits’, ‘Fuzzers’, ‘Generic’, ‘Normal’ and ‘Reconnaissance’ and the rare attack such - ‘Shellcode’ using the Random Forest classifier.

Building a classification model with or without all the features gave similar accuracy, precision and recall scores, from this it can be deduced that the dataset contains a lot of redundant features.

# **7. Business Recommendations & Future Enhancements**

### 7.1 How to improve data collection, processing and model accuracy?

Deep learning is the betterment of the neural network. It became popular in recent years. The current IDS (intrusion detection system) can be improved by using this new technique. The deep learning methods are classified as per their architecture into three types: generative (unsupervised), discriminative (supervised) and hybrid.

To improve efficiency and minimize the training time we need high computing resources which are very costly and require more power. Reinforcement learning (RL) is one of the emerging fields and the research is still going towards attacks detection. Also Deep Reinforcement Learning can be applied as the next step for intrusion detection applications.

Future scopes are provided to help researchers for finding more efficient solutions to detect the attacks. Future directions insist the usage of deep learning and reinforcement learning techniques and Subspace ML for intrusion detection.

### 7.2 Business Recommendations

In recent times, companies have turned to using data analytics to detect fraud. Predictive analytics has expanded the capabilities for fraudulent transaction detection and has experienced a wide adoption among banks and credit card companies in particular.

How companies can use it:

* Financial institutions using it to identify frauds in leasing contracts
* Banks are using it to detect credit card, wire transfers, check frauds
* Insurers are using it to detect fraudulent claims to save the losses
* Healthcare provider can optimize the medical loss ratio by detecting claims frauds