**

SIT742 Modern Data Science

Assignment 2

Team Members:

|  |  |  |  |
| --- | --- | --- | --- |
| Student Name | Yash Ahuja | Priya Singh | Somya Kapoor |
| Student ID | 219608443 | 219592426 | 219339951 |

Date: 26/05/2021

Table of Contents

[**Part I - Data Analytic — Web Log Data** 3](#_Toc69069890)

[**1.** **Data ETL** 3](#_Toc69069891)

[**1.1.** **Load Data** 3](#_Toc69069892)

[**1.2.** **Feature Selection** 3](#_Toc69069893)

[**2.** **Unsupervised learning** 4](#_Toc69069894)

[**3.** **Supervised learning** 5](#_Toc69069895)

[**3.1.** **Data Preparation** 5](#_Toc69069896)

[**3.2.** **Logistic Regression** 6](#_Toc69069897)

[**3.3.** **K-fold Cross Validation** 8](#_Toc69069898)

[**4.** **Association Rule Mining** 10](#_Toc69069899)

[**Part II - Web Crawling** 15](#_Toc69069900)

[**5.** **Crawl the professor Gang Li citation information from 2003 to 2021** 15](#_Toc69069901)

[**6.** **Train Arima to predict the 2018 to 2020 citation** 17](#_Toc69069902)

[**6.1.** **Train Arima Model** 17](#_Toc69069903)

[**6.2.** **Predicting the citation and Calculate the RMSE** 17](#_Toc69069904)

[**6.3.** **Draw the visualization to compare** 19](#_Toc69069905)

[**7.** **Conduct the Grid Search with parameter selection and then predict the 2021 and 2022** 20](#_Toc69069906)

[**7.1.** **Grid Search** 20](#_Toc69069907)

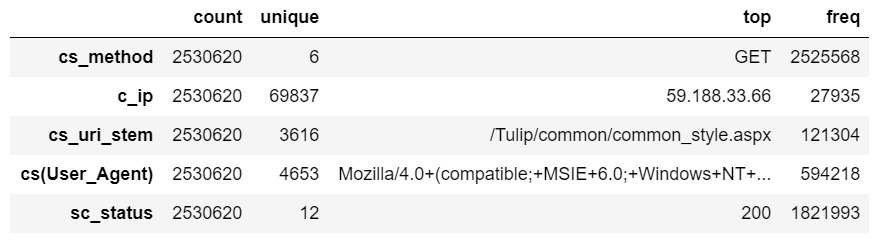
[**7.2.** **Select the best parameter values and Predict for 2021 and 2022** 22](#_Toc69069908)

[**Part III - Self Reflection - Essay** 23](#_Toc69069909)

**Part I - Data Analytic — Web Log Data**

Hotel TULIP a five-star hotel located at Deakin University, and its CIO Dr Bear Guts has asked the Team-SIT742 team to analyse the weblogs files. As an employee for Hotel Tulip, working in the Information Technology Division, it is required to prepare a set of documentation for Team-SIT742 to allow them to understand the data being dealt with. Throughout this report, some source codes are to explore the weblog, which afterwards the information is presented to Dr Bear Guts in the format of a report.

1. **Data ETL**
   1. **Load Data**
2. *The number of requests in weblog\_df is* ***2530620***
   1. **Feature Selection**
3. *Data Description of ml\_df.*

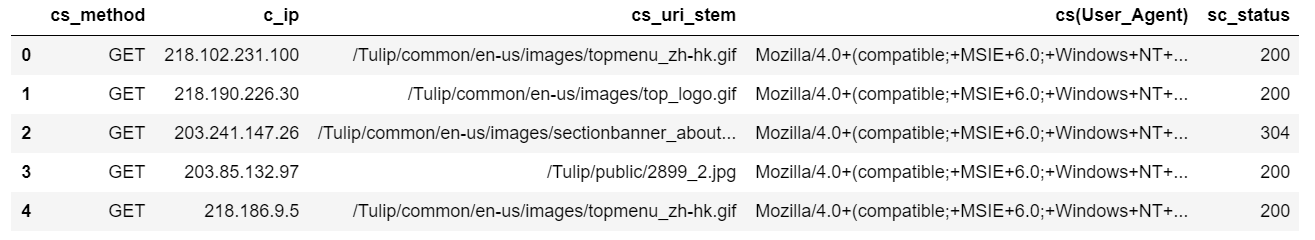


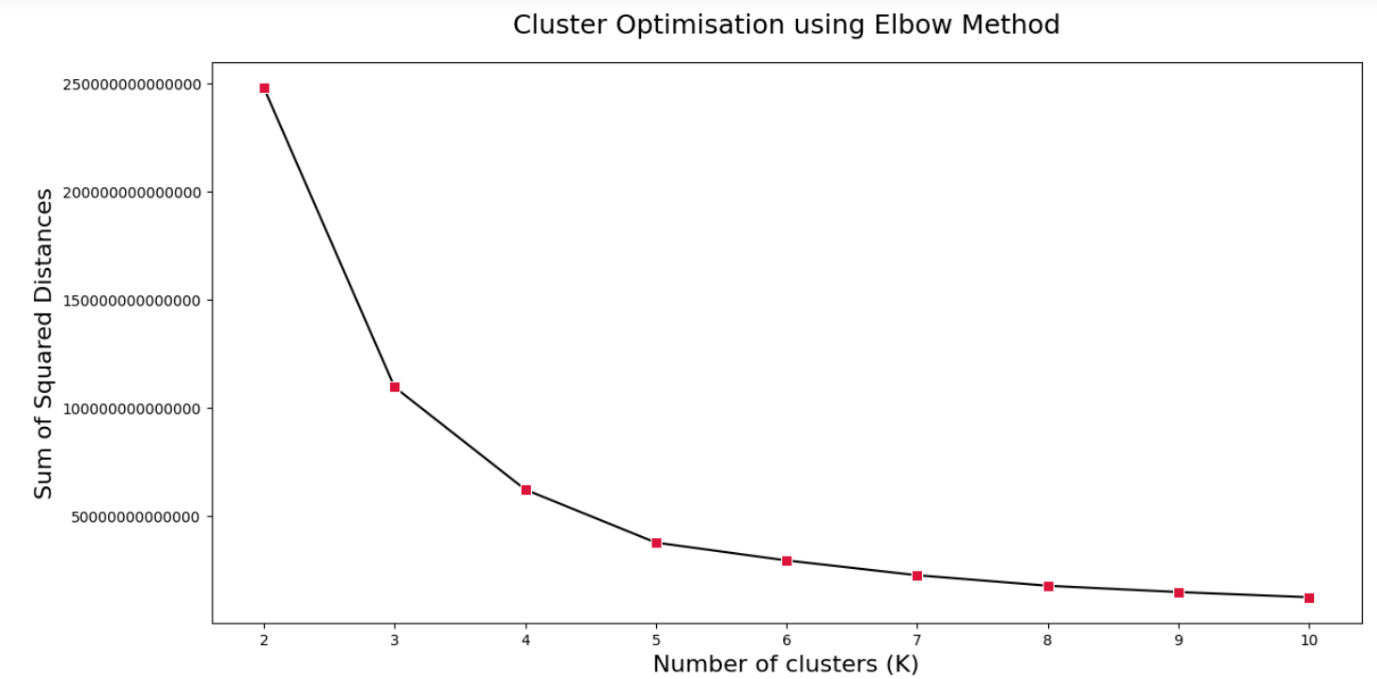
***Count*** denotes the number of total requests(records) for that attribute.

***Unique*** denotes the number of unique requests for that attribute.

***Top*** denotes the example value of the attribute which has highest frequency.

***Frequency*** denotes the number of times the **top** value of the attribute has appeared in total number of requests*.*

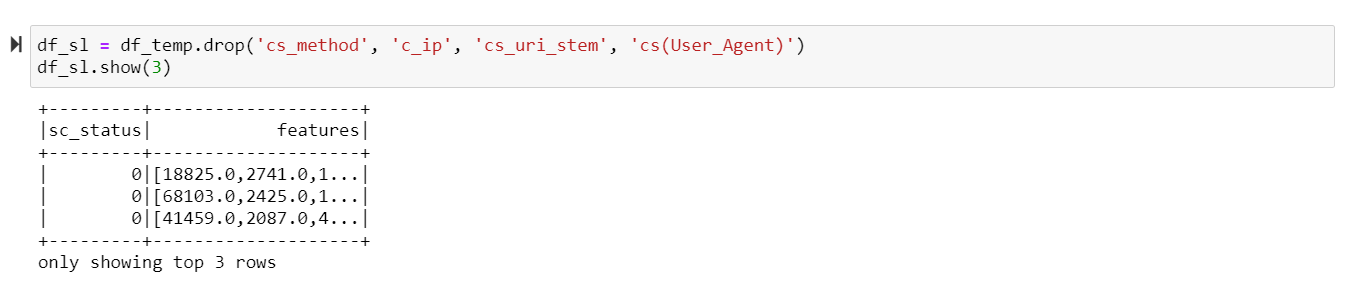
1. *Top 5 rows of ml\_df.*
2. **Unsupervised learning**

Figure 'KMeans' in the elbow plot, with a varying K from 2 to 10.

From the above figure, it is evident that Optimal value of K for this dataset is **5** as it is the smallest value of K where WCSSD is low and increasing the number of clusters will not significantly reduce the WCSSD. A sharp line angle or an 'elbow' on the arm is identified at this point **K=5** implying the best value of K. Hence, the name of the method is elbow method.

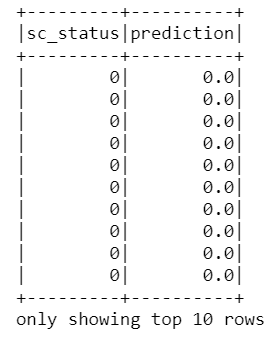
*WCSSD* is Within Cluster Sum of Squared Distances, a key metric for cluster evaluation. A low WCSSD value indicates high similarity within clusters and therefore implying good clustering.

1. **Supervised learning** 
   1. **Data Preparation**

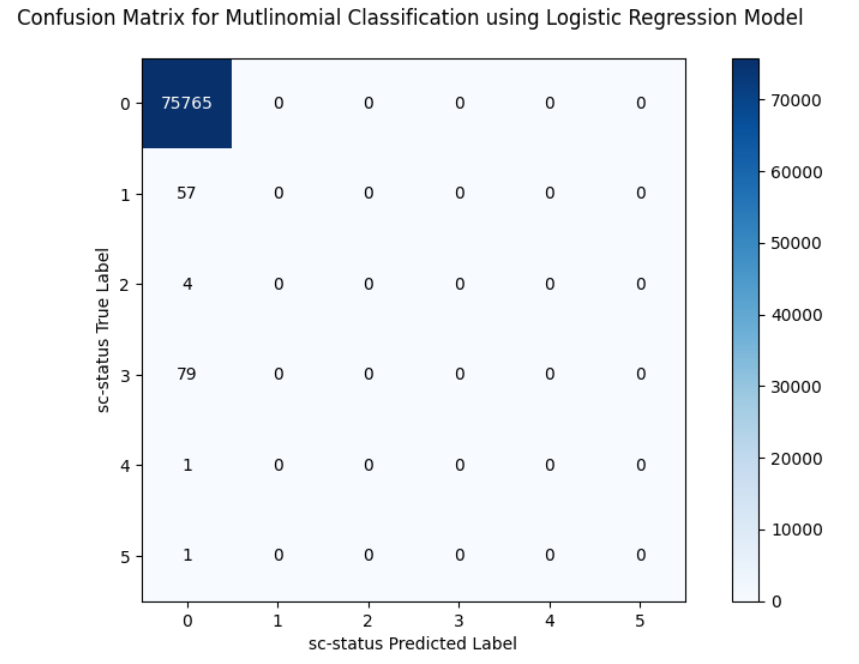


* 1. **Logistic Regression**

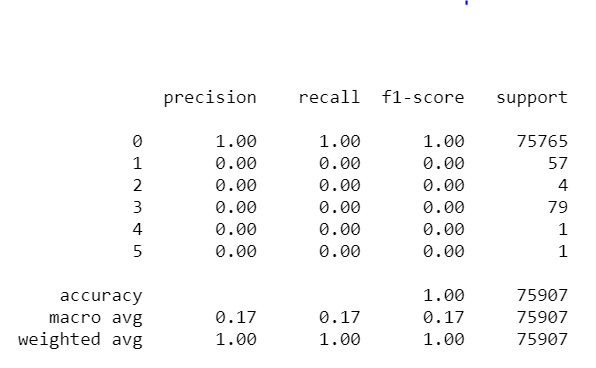
1. *Displaying the classification result using confusion matrix including TP, TN, FP, FN,*



Printing top 10 rows containing observations and predictions of sc\_status leveraging multinomial classification on ml\_df using logistic regression model:



1. *Displaying the classification result using Precision, Recall and F1 score.*



It shows the logistic regression model classification result using classification performance metrices such as precision, recall, f1-score.

**Recall:** True Positive / (True Positive + False Negative).

Out of all the items that are truly positive, how many were correctly classified as positive. Or simply, how many positive items were 'recalled' from the dataset.

**Precision:** True Positive / (True Positive + False Positive).

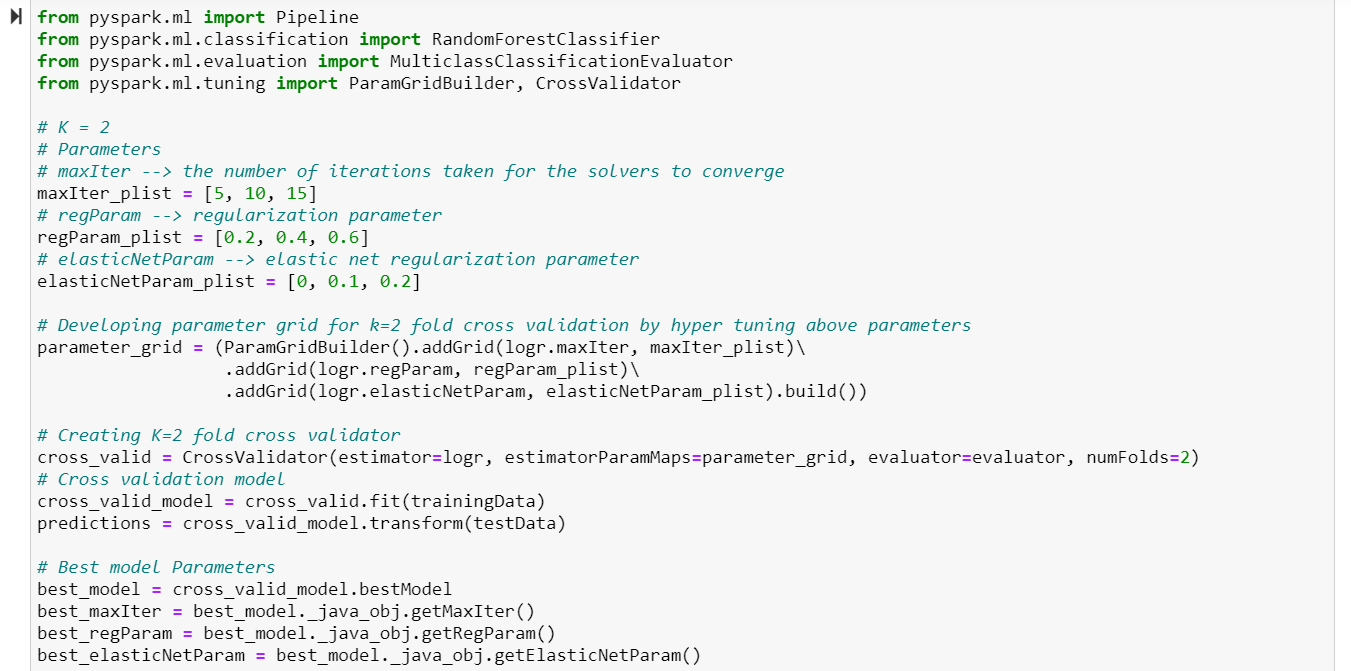
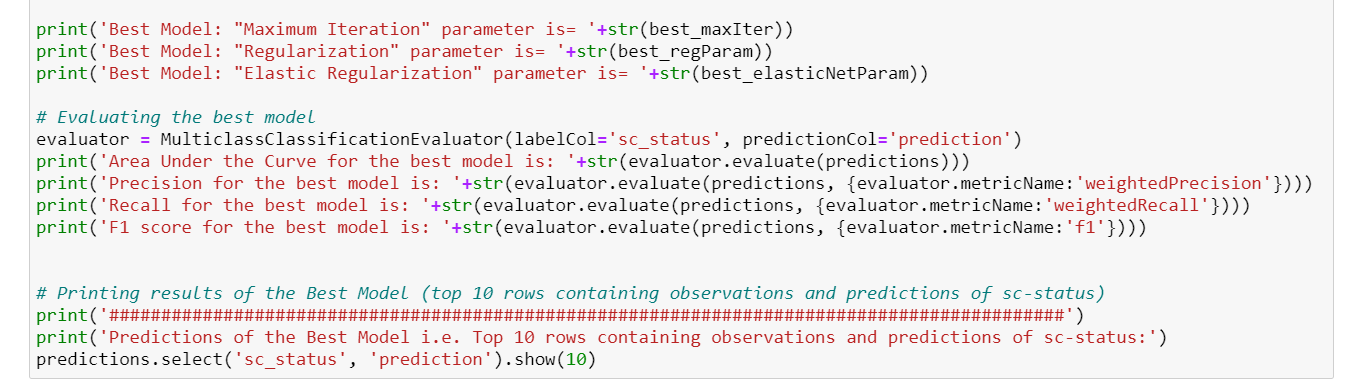
Out of all the items labeled as positive, how many truly belong to the positive class.

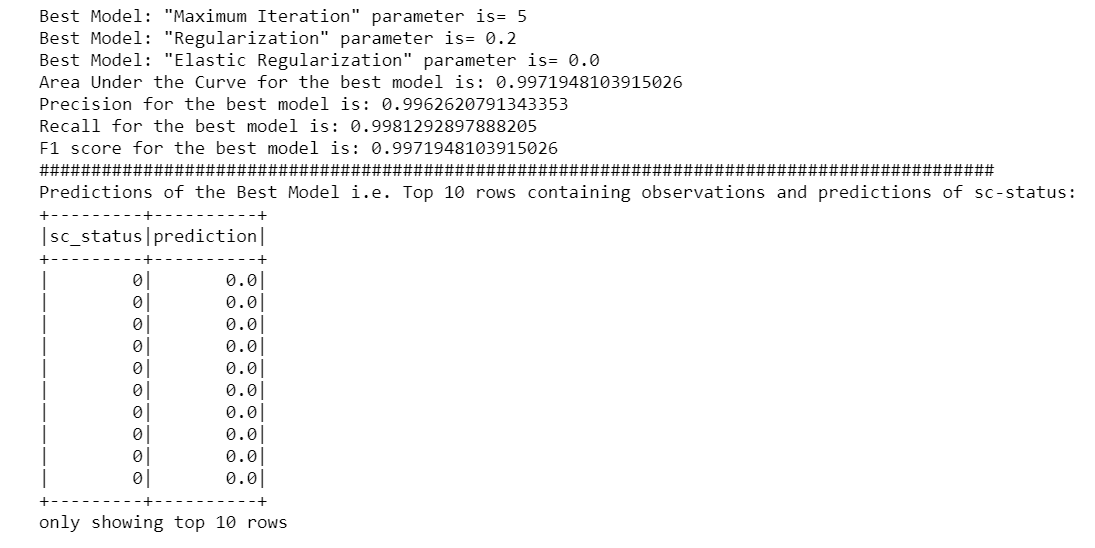
**F1 Score:** 2\*((Precision\*Recall)/(Precision + Recall)).

It is a measure to illustrate the balance between the Recall and the Precision. Therefore, also known as F Measure or F Score.

* 1. **K-fold Cross Validation**

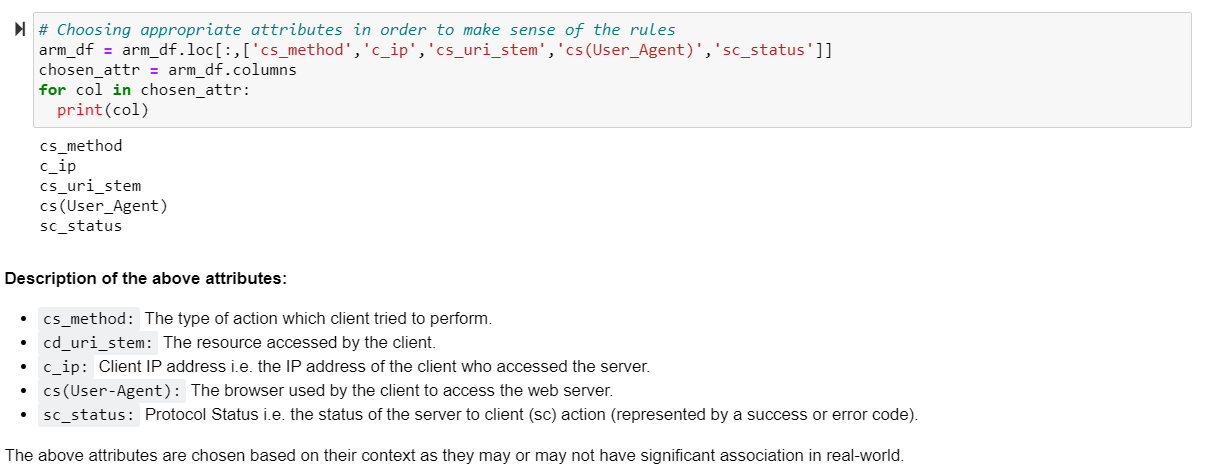
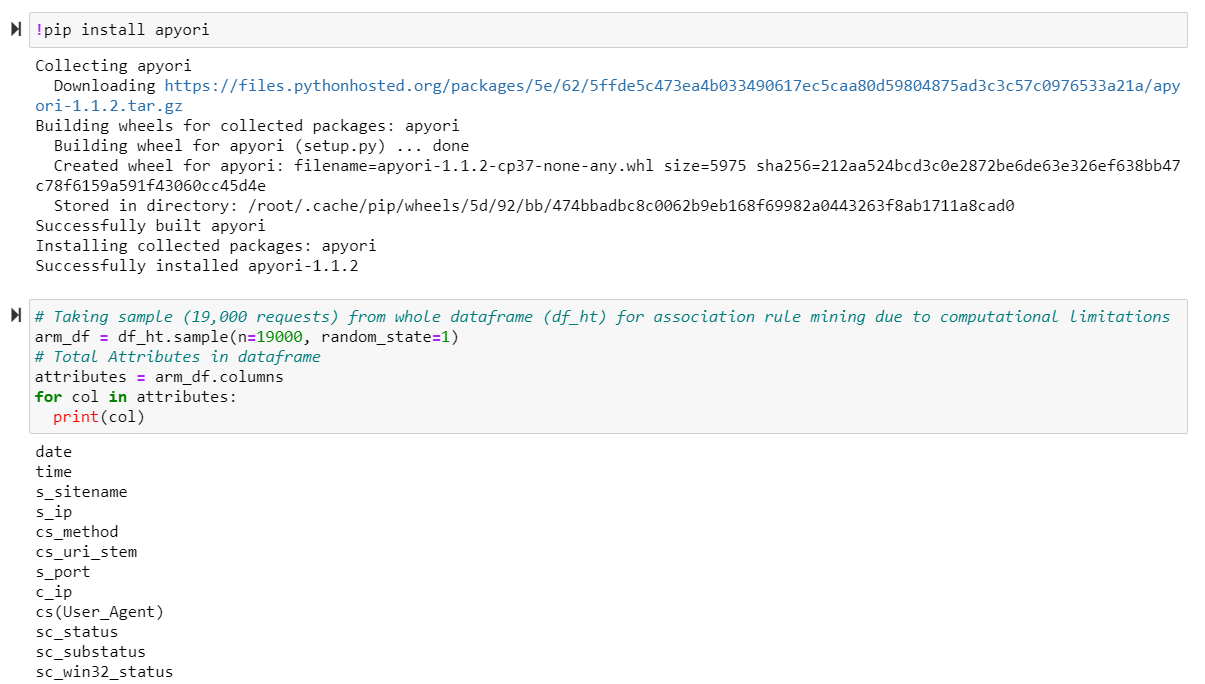
1. *Our code design and running results*

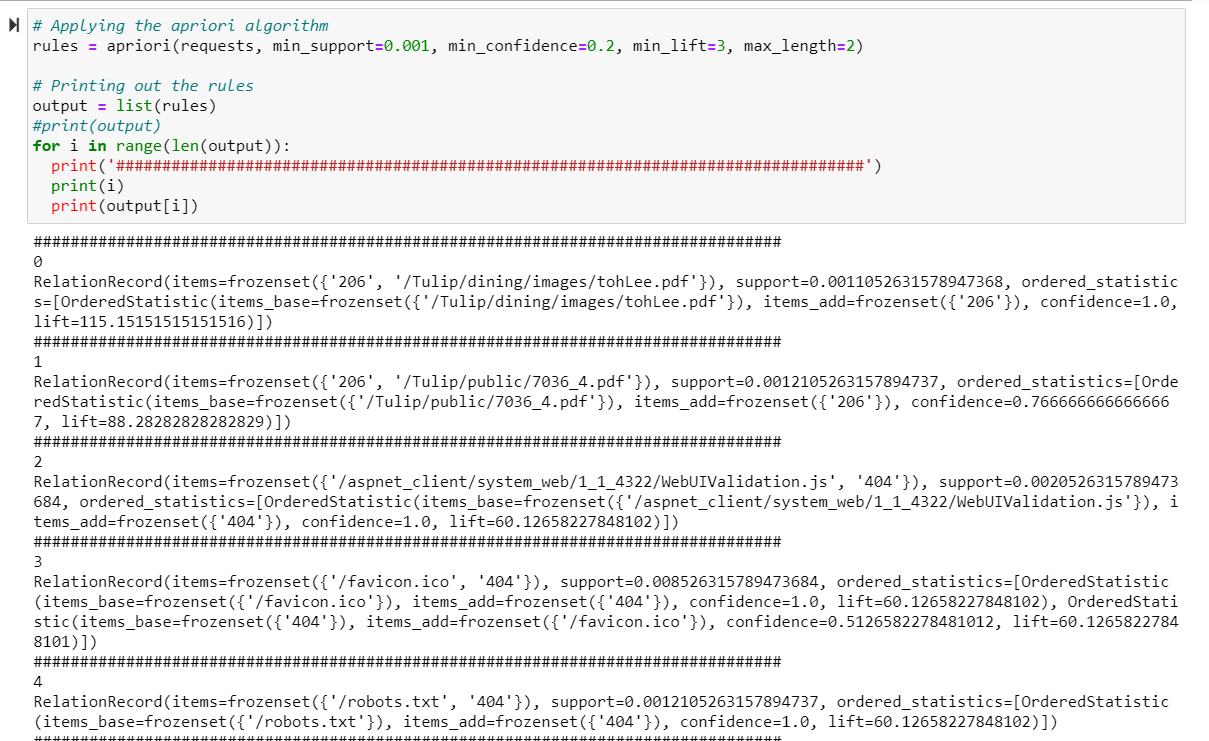




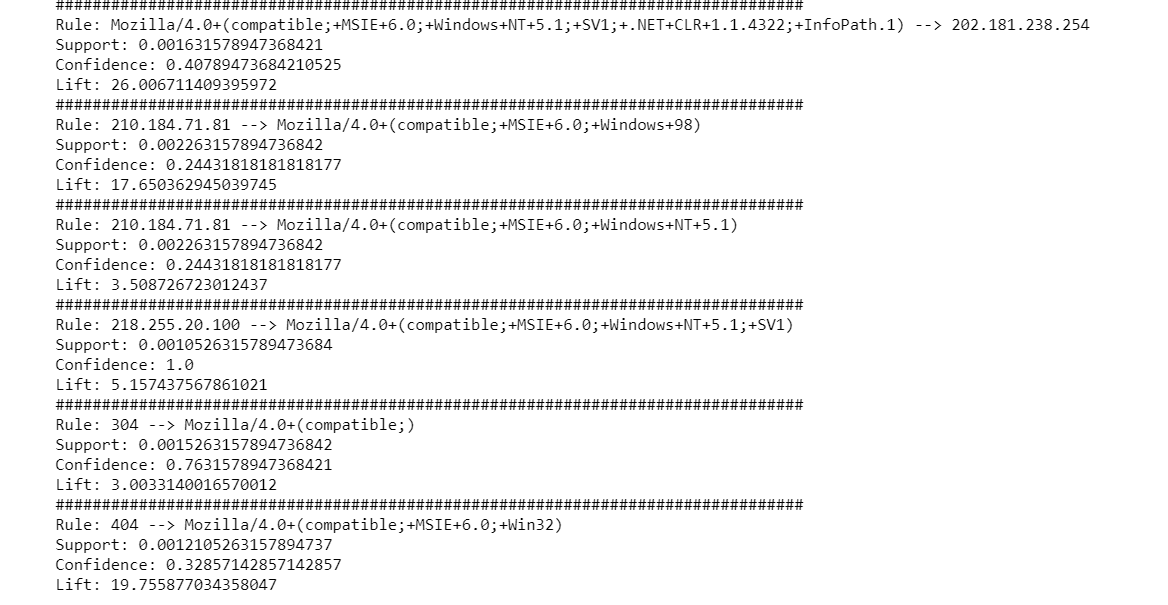
1. *Our findings on hyper-parameters based on this cross-validation results (Best results).*

It is evident that the parameters used for training i.e. maxIter = 5, regParam= 0.2, elasticNetParam=0 are found to be the ones which have the best model performance after performing K=2 fold Cross Validation and that is why Area Under the Curve, Precision, Recall and F1 score metrics are same for training model and cross validation model.

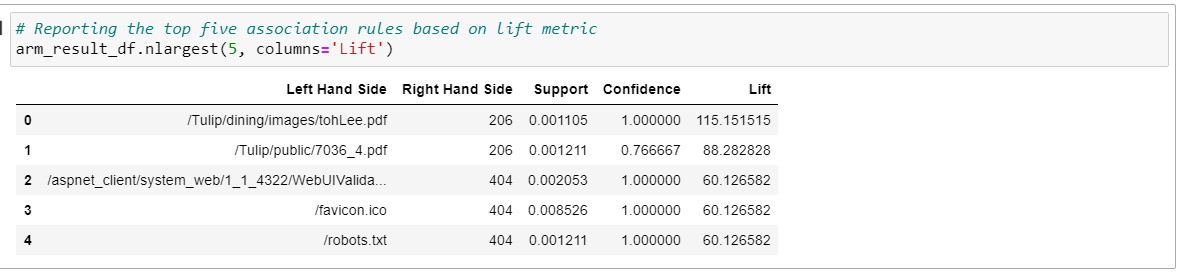
1. **Association Rule Mining**
2. *Our code design and running results, threshold (parameter) for support and confidence*







1. *Our findings on association rule mining results*



**It is evident from the above output of top five association rules that resource used by the client i.e. cs-uri-stem attribute has significant association with the protocol status (the status of the server to client action, usually represented by a code) i.e. sc-status attribute which makes sense in the real-world.**

**Part II - Web Crawling**

In 2021, to better introduce and understand the research works on the professors, Deakin university wants to perform the citation prediction on individual professor level. You are required to implement a web crawler to crawl the citation information for Gang Li from 2003 to 2021 (start at 2003 and end at 2021), and conduct several prediction coding tasks. You will need to make sure that the web crawling code and prediction code meets the requirements. You are free to use any Python package for Web crawling and prediction by finishing below tasks.

1. **Crawl the Gang Li citation information from 2003 to 2021**

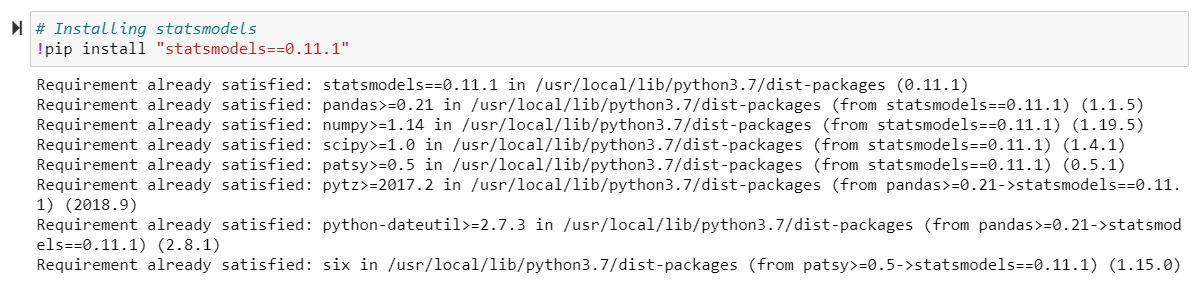
*Screenshot of our code for generating citation2003-2021.csv.*

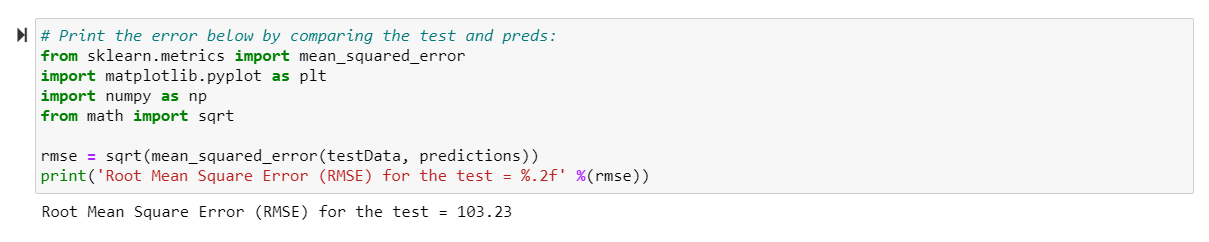


1. **Train Arima to predict the 2018 to 2020 citation** 
   1. **Train Arima Model**

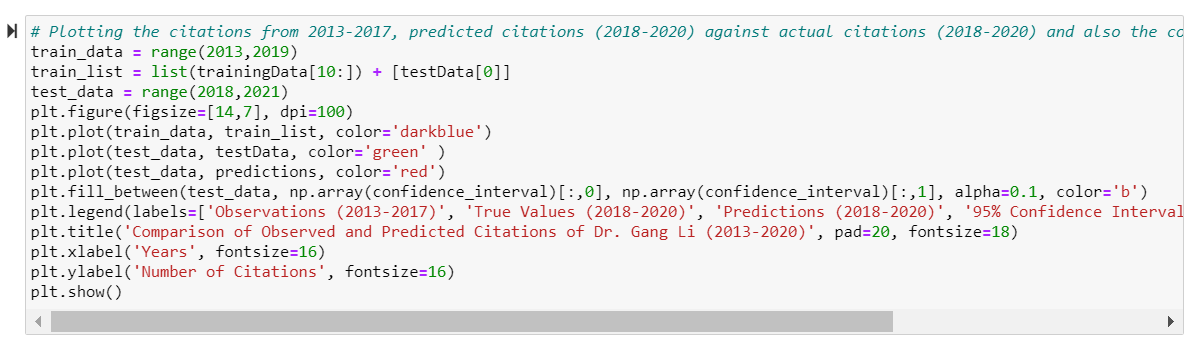
*Screenshot of our code for Arima Training.*

* 1. **Predicting the citation and Calculate the RMSE**

*Screenshot of our code for predicting and display RMSE (root mean square error).*



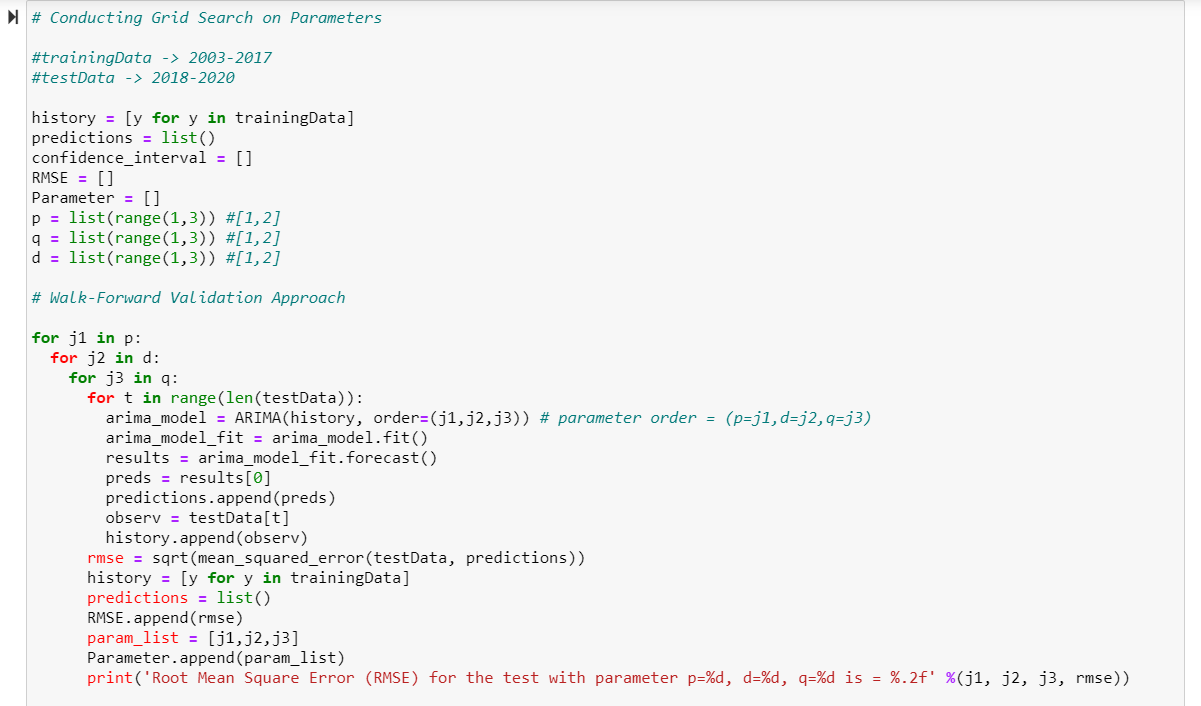
* 1. **Draw the visualization to compare**

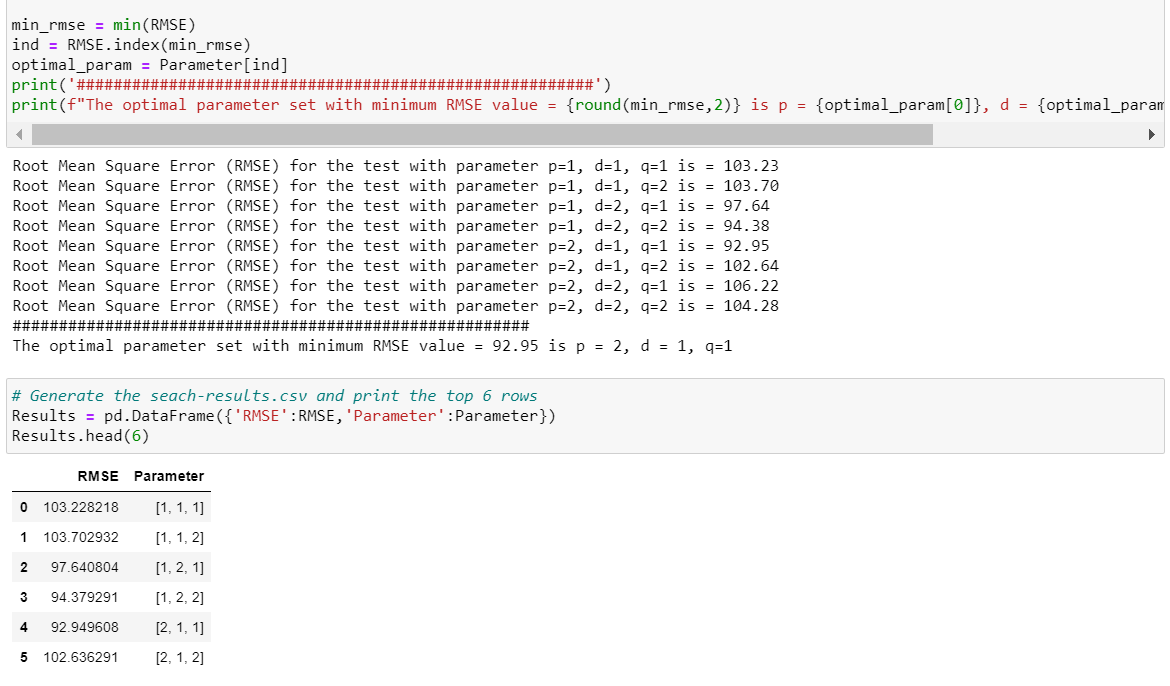
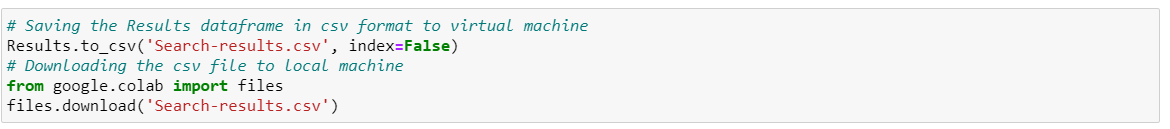
*Screenshot of our line plot, along with its code.*

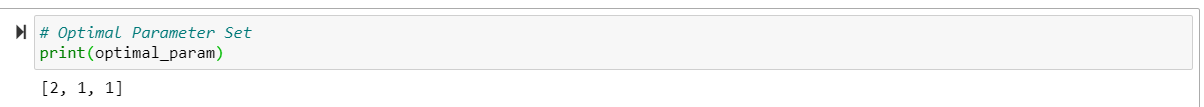
Chart, line chart

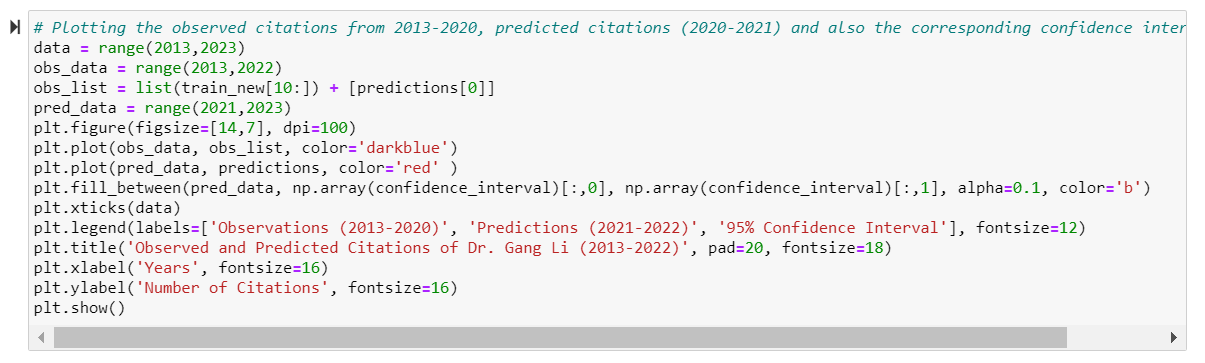
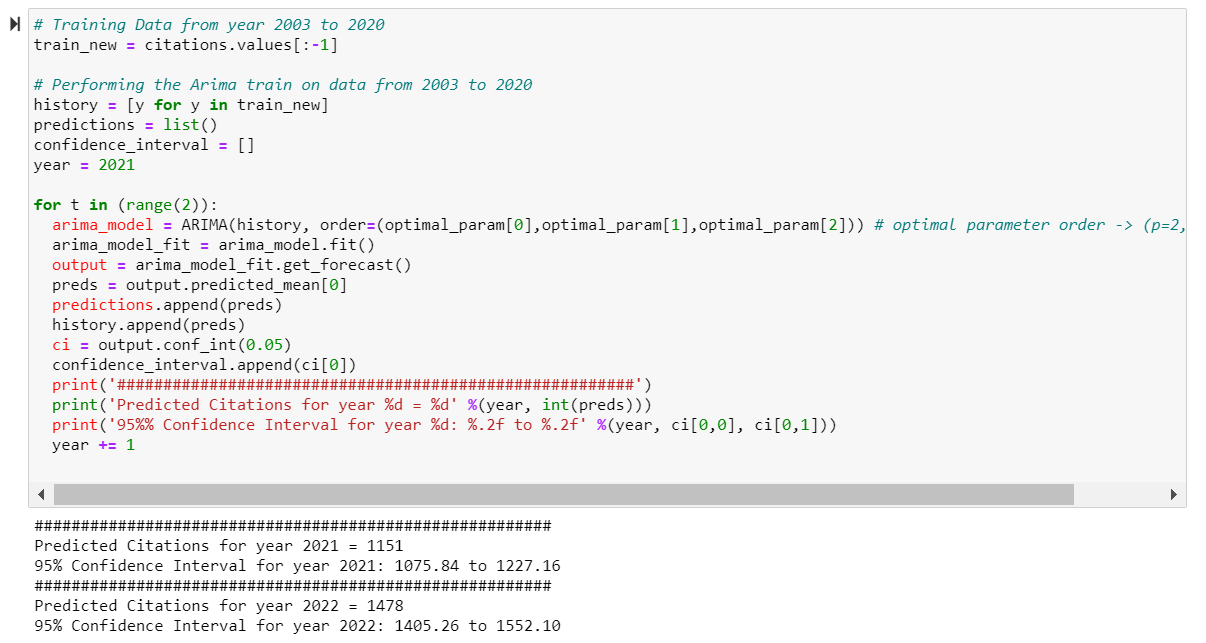
Description automatically generated

1. **Parameter selection and Year 2021 and 2022 Prediction**
   1. **Grid Search**

*Screenshot of our code for grid search.*



* 1. **Select the best parameter values and Predict for 2021 and 2022**
     1. *Displaying the best parameter values*
     2. *Line plot with training data from 2013 to 2020, the predictions together with the confidence interval.*



95% Confidence Interval for year 2021*:* ***1075.84 to 1227.16***

95% Confidence Interval for year 2022*:* ***1405.26 to 1552.10***

*Chart, line chart

Description automatically generated*

**Part III - Self Reflection - Essay**

**1. What are the Python packages that you find useful in manipulating and analysing Big data? You can briefly analyse their advantages and disadvantages:**

***Pandas*:** Pandas is open-source python package. Pandas is used for data analysis and data structures and operations for manipulating numerical tables and time series.

1. **Advantages**:
   * 1. Extensive features set.
     2. Makes data flexible and customizable.
2. **Disadvantages:**
   * 1. Steep Learning curve.
     2. Complex syntax.

***Numpy:*** It is used for scientific and numerical computation in Python. It allows working on large datasets and storage.

1. **Advantages:**
   * 1. Takes less memory space and provide better runtime speed.
     2. Support Vectorized operations
2. **Disadvantages:**
   * 1. Requires contagious allocation of memory.

***Scipy:*** Scipy is scientific computation library in Python. It provides utility functions for optimization, stats, and signal processing.

1. **Advantages:**
   * 1. SciPy contains various versions of the linear algebra modules and other numerical algorithms.
2. **Disadvantages:**
   * 1. Scipy is lower in computation than Numpy.

**Matplotlib:** It is python library used for 2d graph such as bar graph,histogram, pie chart for plotting.

1. **Advantages:**
   * 1. Easy to use for basic plots such as histogram, pie chart etc.
2. **Disadvantages:**
   * 1. Reliant on other packages to work such as Numpy.

**2. What are the Big data platforms that can help storing, retrieving, and analysing the big data? What are their advantages and disadvantages?**

Some of the Big data platforms that can store, retrieve, and analyse the data are described as follows:

**Apache Hadoop:** Apache Hadoop is java-based software used to store large amount of data in cluster and run them parallelly across several nodes. The storage system of Hadoop is known as HDFS – Hadoop Database File System. It divides the data into small sets and process them in various nodes.

**Advantages:** Scalable, flexible, fault-tolerant, robust

**Disadvantages:** Not ideal for smaller datasets, security issues

**NoSql:** NoSQL is used to deal with unstructured datasets. These datasets are stored with no schema; however, it gives better performance in storing huge amount of data and analysing them.

**Advantages:** Low cost (open source), easily scalable, can store unstructured data (such as text documents)

**Disadvantages:** Lack of standardization, insignificant community support

**Hive:** Hive is also one database management system in BigData. It provides query options like SOL query and primarily used for data mining purposes.

**Advantages:** Data may be queries by several users at the same time, easy implementation of ETL jobs

**Disadvantages:** Used extensively for online analytical processing systems rather than online transaction processing systems.

**Datawrapper:** Data wrapper is open-source software for data visualization to generate charts and analysis on the datasets. It is very device friendly, fast, and interactive tool to depict the visualization of large datasets.

**3. Compare and contrast the Python data analytical packages and their Spark packages.**

Two spark packages are:

1. Apache Spark
2. PySpark

**Apache Spark:** Apache Spark is computing engine; it contains set of libraries which helps in parallel computing on computer clusters. It is open source. It is used for large scale data processing

**Pyspark**: Pyspark is python API, it is collaboration of apache spark and python, It is also open source and used for cluster computing framework. It is created and licenced under Apache Spark Foundation. It used library called Py4j, an API written in Python. It is especially used in big data.

**Python data analytical packages:** Python has standard library that supports a wide variety of functionalities like databases, automation, text processing, scientific computing. Python packages are generally easy to use, high performance and used in AI. Some of the python packages are Pandas, Numpy, SciPy, Seaborn, TensorFlow etc.

**4. What are your opinions on the privacy issues in the Big data era? Any example to further illustrate the risks?**

There are many benefits from Big Data analytics, but also have a massive potential for exposure that could result in anything. Mentioned below are some of the ways to prevent the risks in big data.

Some ways to prevent the exposure risk in big data are:

1. **Govern Your Data Access**

We live in an age where a mail truck full of bank account statements getting hi-jacked is unlikely, but a data warehouse full of social security numbers or intellectual property data getting hacked into is very prevalent. To combat these faceless culprits, you should take these steps to improve your understanding of who has access to your treasured data:

* Assess current access requirements and processes, key systems, or applications containing large volumes of data within your company environment. This is not just limited to internal employees, but also any third parties who may be receiving or using the data.
* Define baseline access requirements for these key systems or applications containing large datasets. This includes who should have access and what business justification should warrant access.
* Establish access controls for these big data stores such as strong authentication, and approvals from data owners before granting access. Or, reduce the number of users by appointing a central “librarian” to control access to data stores.
* Perform ongoing monitoring of user access against the baseline requirements. This will help you to proactively identify deviations from normal access and quickly address insufficient controls.

1. **Limit the Data Use, Collection and Storage**

Organizations often collect and hoard information that is not really required or even used, perhaps thinking there might be a need for it someday. As defined in the AICPA Generally Accepted Privacy Principles (GAPP), basic privacy principles such as “collection limited to identified purpose” and “use, retention, and disposal” urge organizations to:

* Review current data residing in big data stores and determine the business need for collecting, using, and storing them. If you are not using the data for legitimate business reasons, you should stop collecting it, as it only introduces more risk exposure.
* Understand the upstream and downstream data flow of your big data store. Sometimes the information is “fed” to us by a third party. If you want to stop receiving unnecessary data from the source, you need to first understand where it comes from.
* Follow the retention policy or schedule and diligently archive data into an approved archival solution. This will not only help you to reduce data related breaches and comply with retention requirements but will also improve capacity management of your networks and systems.

1. **Leverage Technologies**

Additionally, relying on technologies to help us secure big data is imperative. Protecting your data through end-to-end encryption or tokenization will help to minimize data from being understood by unauthorized people. However, consider these basic impacts of technology implementation on general accessibility and availability of data:

* Performance impact on data access. A few seconds may not seem to be a deal breaker, but if there are multiple databases and multiple applications across multiple platforms involved, then the sum of that lead time might be unacceptable to the end users or critical business processes.
* Business process impact. Utilizing tokenization to substitute a sensitive data element that cannot be mathematically reversed is a popular way to reduce Payment Card Information (PCI) Data Security Standard (DSS) scope. With the use of a token instead of payment card data, it might alter some of the business processes handling of payment card data.
* Complexity associated with key management. Without effective key management policies and processes, the encryption is no more secure than handing your keys over to the inmates.

Big data does not always come with supersized risks. You can reduce big data breaches by defining access requirements; limiting the collection, use, or storage of data to only support your business need; and applying technical controls to protect data from intruders.

**5. What are the methods you think could help to solve the privacy issues on big data? Please list any successful implemented method.**

Below are some privacy issues and how to solve them on Big data:

**Data privacy best practices for big data**

There are certain strategies organizations can use to protect big data. Several of the best practices for maintaining the privacy of big data include:

1. **Employ real-time monitoring.**

Since a privacy issue can happen at any moment, organizations should find a solution that monitors data in real-time. This way, they will be aware of a problem as soon as it happens and can take appropriate steps to resolve it right away.

1. **Implement homomorphic encryption.**

Homomorphic encryption is a form of encryption that allows users to compute data without decrypting it first. This form of encryption should be implemented to store and process information in the cloud to prevent organizations from revealing private information to outside vendors.

1. **Avoid collecting too much data.**

Only the data that is necessary should be collected. An organization may not need the Social Security numbers of their customers; customer login usernames and passwords may only be necessary. Organizations should consider deleting any personal information that is not needed to best protect customer data privacy.

1. **Prevent internal threats.**

Organizations are also exposed to internal privacy risks from angry or simply uninformed employees. Therefore, it is essential to educate all employees on best practices for ensuring data privacy like changing passwords frequently and logging off unused computers.

**6. Any other thoughts about data science, or suggestions to future students (or teaching team) about this unit.**

After completing the SIT742 Modern Data Science, through experiences we can suggest various things for future students. Some of them are:

Students who do not have python background, may find it difficult to understand advanced machine learning libraries and implantation of the concepts of data science. Thus, we recommend prerequisite of any python related unit or Machine learning unit as a mandatory unit for this unit.

Also, unit can arrange some extra online help hub sessions as refresher for the unit.

For teaching team, we suggest increasing the tutorial time and include the time for assessments related doubts. This will not only improve the understanding of assessments among the students but also improve the performance in the same. We highly recommend the teaching team to include every topic from the assessment to be covered in the tutorials, for example web crawling and other such topics used in assessment 2. Proper guidance can be provided through links and study material to study and understand those topics instead of doing research online.