Predicting Computer Access Report

Student:

Kiran Dere: 13

Shubham Jakhete: 26

Pooja Kadam: 28

Teacher:

Mrs. Yogini Kulkarni

Course:

M.SC Computer Science Part 1

# Index

|  |  |
| --- | --- |
|  | Title |
| 01 | What is Data Analysis |
| 02 | Data Analysis Working  Data Requirement  Data Processing  Univariate Data Analysis  Modeling and Algorithms |
| 03 | Literature Study  Amazon Employee Access Challenge |
| 04 | Feasibility Study  Economic Feasibility  Technical Feasibility  Social Feasibility |
| 05 | System Requirement Specification |
| 06 | UML Design  Use-Case Diagram  Class Diagram  Sequence Diagram  Artifact Diagram |
| 07 | Data Dictionary |
| 08 | Implementation  Importing Packages  Univariate Data Analysis  Data Transformation  Modeling and Algorithms |
| 09 | Code Screenshots and Explanation |
| 10 | Testing  Blackbox Testing  Test Cases |
| 11 | Conclusion |
| 12 | References |

# What is Data Analysis?

* Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.
* Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.
* Data mining is a particular data analysis technique that focuses on modelling and knowledge discovery for predictive rather than purely descriptive purposes, while business intelligence covers data analysis that relies heavily on aggregation, focusing on business information.
* Predictive analytics focuses on application of statistical models for predictive forecasting or classification.
* The term data analysis is sometimes used as a synonym for data modelling.

# Data Analysis Working

Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories. There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases.

Procedure:

## Data Requirement:

* The data is necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis or customers (who will use the finished product of the analysis).
* The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people).

## Data Processing:

* The phases of the intelligence cycle used to convert raw information into actionable intelligence or knowledge are conceptually similar to the phases in data analysis.
* Data initially obtained must be processed or organized for analysis.

## Univariate Data Analysis:

* This type of data consists of only one variable.
* The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
* It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

## Data Transformation:

* Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system.
* Data transformation is a component of most [data integration](https://www.talend.com/resources/what-is-data-integration/) and [data management](https://www.talend.com/resources/what-is-data-management/) tasks, such as data wrangling and [data warehousing](https://www.talend.com/resources/what-is-data-warehouse/).

## Modeling and Algorithms:

* Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation.
* In general terms, models may be developed to evaluate a particular variable in the data based on other variable(s) in the data, with some residual error depending on model accuracy (i.e., Data = Model + Error).
* Inferential statistics includes techniques to measure relationships between particular variables.
* For example, regression analysis may be used to model whether a change in advertising (independent variable X) explains the variation in sales (dependent variable Y). In mathematical terms, Y (sales) is a function of X (advertising).
* It may be described as Y = aX + b + error, where the model is designed such that a and b minimize the error when the model predicts Y for a given range of values of X.
* Analysts may attempt to build models that are descriptive of the data to simplify analysis and communicate results.

# Literature Study

## Amazon Employee Access Challenge:

* When an employee at any company starts work, they first need to obtain the computer access necessary to fulfil their role.
* This access may allow an employee to read/manipulate resources through various applications or web portals.
* We are building a model, that will determine an employee's access needs, such that manual access transactions (grants and revokes) are minimized as the employee's attributes change over time.
* The model will take an employee's role information and a resource code and will return whether or not access should be granted.
* The objective of new system is to build a model, learned using historical data, that will determine an employee's access needs, such that manual access transactions (grants and revokes) are minimized as the employee's attributes change over time.
* The model will take an employee's role information and a resource code and will return whether or not access should be granted. These auto-access models seek to minimize the human involvement required to grant or revoke employee access.

# Feasibility Study

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are:

* Economic Feasibility
* Technical Feasibility
* Social Feasibility

## Economic Feasibility:

* This study is carried out to check the economic impact that the system will have on the organization.
* The amount of fund that the company can pour into the research and development of the system is limited.
* The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

## Technical Feasibility:

* This study is carried out to check the technical feasibility, that is, the technical requirements of the system.
* Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources.
* This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

## Social Feasibility:

* The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently.
* The user must not feel threatened by the system, instead must accept it as a necessity.
* The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it.
* His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

Existing System:

* It is often the case that employees figure out the access they need as they encounter roadblocks during their daily work.
* A knowledgeable supervisor then takes time to manually grant the needed access in order to overcome access obstacles.
* As employees move throughout a company, this access discovery /recovery cycle wastes a nontrivial amount of time and money.

Proposed System:

* We proposed an improved analytics system, learned using historical data, that will determine an employee's access needs, such that manual access transactions (grants and revokes) are minimized as the employee's attributes change over time.
* The model will take an employee's role information and a resource code and will return whether or not access should be granted. These auto-access models seek to minimize the human involvement required to grant or revoke employee access.

# System Requirement Specifications

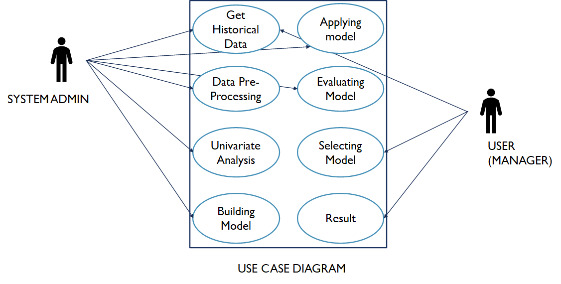
* **Functional Requirements**
  + Storage and conversion of Excel and CSV files
  + Visualization of data
  + Python packages
  + Machine Learning packages
* **Non Functional Requirements** 
  + Quality
  + Response Time
  + Security
  + Supportability

**Operating Specifications**

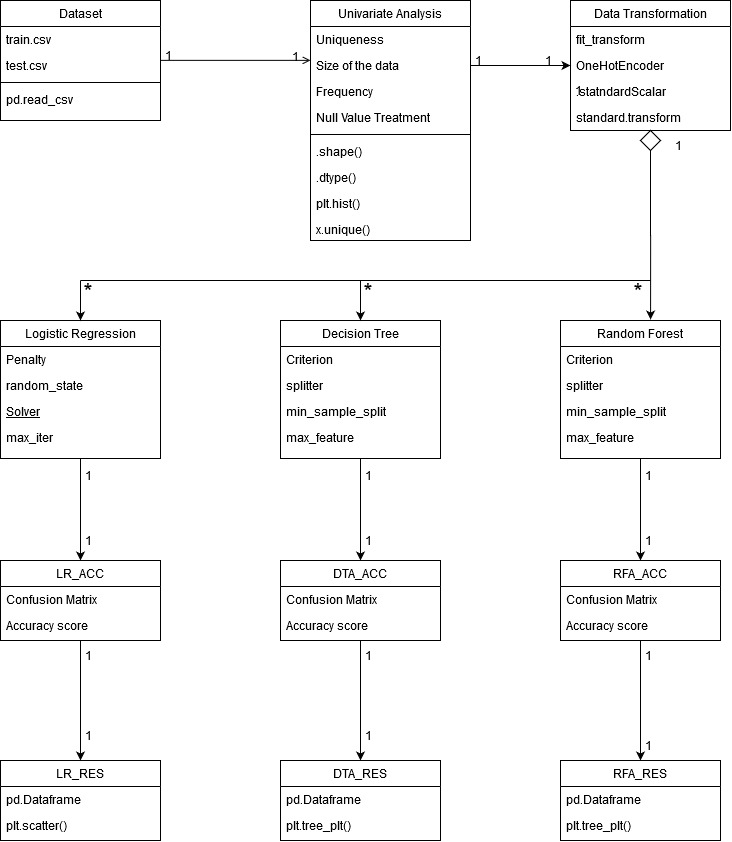
* Minimum Hardware Configuration
  + System : 2.4GHz dual-core processor
  + Storage : 5GB of free disk space
  + Graphics : OpenGL support for GPU hardware for visualization
  + Peripherals : Mouse, keyboard and HD monitor for navigating UI
  + RAM : 8 GB
* Minimum Software Configuration
  + Operating system : Windows, Linux, UNIX or MacOS with Anaconda 3.0
  + Coding Language : Python
  + Coding Ground : jupyterLab
  + File System : Excel File

# UML Design

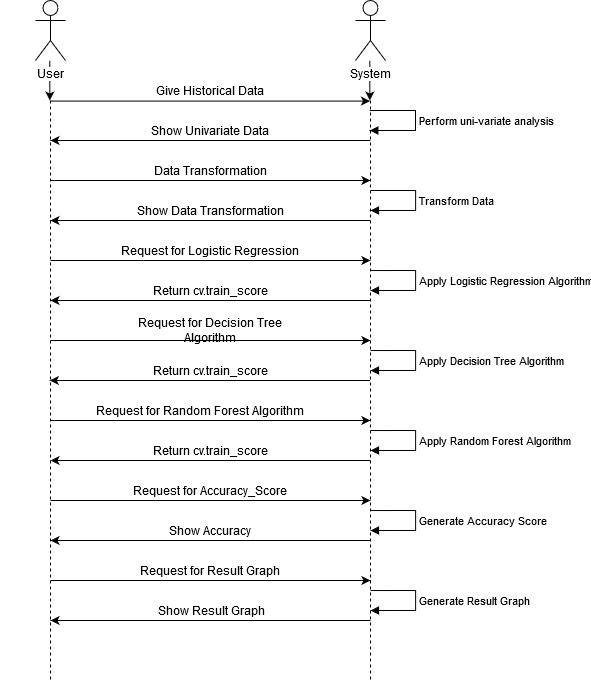
Use Case Diagram:

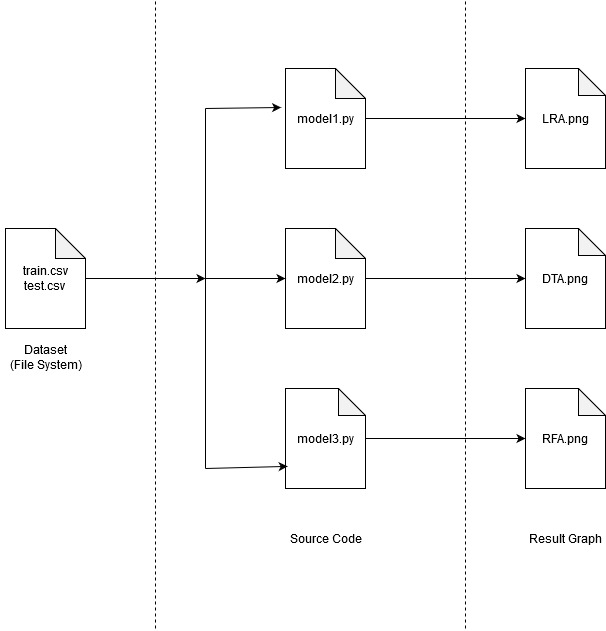


Class Diagram:



Sequence Diagram:



Artifact Diagram: 

# Data Dictionary

|  |  |
| --- | --- |
| **COLUMN NAME** | **DESCRIPTIONS** |
| Action | ACTION is 1 if the resource was approved, 0 if the resource was not approved |
| RESOURCES | An ID for each resource |
| MGR\_ID | The EMPLOYEE ID of the manager of the current EMPLOYEE ID record; an employee may have only one manager at a time |
| ROLE\_ROLLUP\_1 | Company role grouping category id 1 (e.g. US Engineer) |
| ROLE\_ROLLUP\_2 | Company role grouping category id 2 (e.g. US Retail) |
| ROLE\_DEPTNAME | Company role department description (e.g. Retail) |
| ROLE\_TITLE | Company role business title description (e.g. Senior Eng. Retail Manager) |
| ROLE\_FAMILY\_DESC | Company role family extended description (e.g. Retail Manager, Soft. Engineer) |
| ROLE\_FAMILY | Company role family description (e.g. Retail Manager) |
| ROLE\_CODE | Company role code; this code is unique to each role (e.g. Manager) |

# Implementation:

Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories. There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases.

Methodology:

## Importing Packages:

* The necessary packages are imported for data analysis processing and also for implementing algorithms.
* Pandas
  + Pandas is an opensource library that allows to you perform data manipulation in Python.
  + Pandas library is built on top of NumPy, meaning Pandas needs NumPy to operate.
  + Pandas provide an easy way to create, manipulate and wrangle the data. Pandas is also an elegant solution for time series data.
* NumPy
  + NumPy is an open source library available in Python that aids in mathematical, scientific, engineering, and data science programming.
  + NumPy is an incredible library to perform mathematical and statistical operations.
  + It works perfectly well for multi-dimensional arrays and matrices multiplication.
* Seaborn
  + Seaborn is a library for making statistical graphics in Python.
  + It is built on top of Matplotlib and closely integrated with Pandas data structures.
  + Seaborn aims to make visualization a central part of exploring and understanding data.
  + Its dataset-oriented plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.
* Matplotlib
  + Matplotlib is an amazing visualization library in Python for 2D plots of arrays.
  + Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.
  + One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals.
  + Matplotlib consists of several plots like line, bar, scatter, histogram etc.
* Warnings
  + Warning messages are typically issued in situations where it is useful to alert the user of some condition in a program, where that condition (normally) doesn’t warrant raising an exception and terminating the program.
  + For example, one might want to issue a warning when a program uses an obsolete module.
  + The warnings filter controls whether warnings are ignored, displayed, or turned into errors (raising an exception).
  + The warnings filter maintains an ordered list of filter specifications; any specific warning is matched against each filter specification in the list in turn until a match is found; the filter determines the disposition of the match.

## Univariate Data Analysis:

* This type of data consists of only one variable.
* The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
* The Steps in univariate data analysis include:
  + Identification of variables and data types
  + Analyzing the basic metrics
  + Non-Graphical Univariate Analysis
  + Graphical Univariate Analysis
  + Outlier treatment

## Data Transformation:

* Data Transformation is the process of converting data from one format or structure into another format or structure. It is a fundamental aspect of most data integration and data management tasks such as data wrangling, data warehousing, data integration and application integration.
* Data transformation can be simple or complex based on the required changes to the data between the source (initial) data and the target (final) data.
* Data transformation is typically performed via a mixture of manual and automated steps.
* Tools and technologies used for data transformation can vary widely based on the format, structure, complexity, and volume of the data being transformed.
* Categorical data is transformed using the most straight-forward approach one-hot-encoding using OneHotEncoder transformation from scikit-learn package.
* One-Hot-Encoding
  + For categorical variables where no such ordinal relationship exists, the integer encoding is not enough.
  + In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).
  + In this case, a one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.
* Standard Scalar
  + sklearnits main scaler, the StandardScaler, uses a strict definition of standardization to standardize data. It purely centers the data by using the following formula, where *u* is the mean and *s* is the standard deviation.

x\_scaled = (x - u)/s

## Modeling and Algorithm:

* Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation.
* First Model – Logistic Regression
  + Logistic Regression (also called Logit Regression) is commonly used to estimate the probability that an instance belongs to a particular class.
  + If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (called the positive class, labeled “1”), or else it predicts that it does not (i.e., it belongs to the negative class, labeled “0”). This makes it a binary classifier.
  + A Logistic Regression model computes a weighted sum of the input features (plus a bias term), but instead of outputting the result directly like the Linear Regression model does, it outputs the logistic of this result.
  + Logistic Regression model estimated probability (vectorized form)

p̂ = hθ(x) = σ (θT ・x)

* + The logistic—also called the *logit*, noted *σ* (・)— is a *sigmoid function* (i.e., S-shaped) that outputs a number between 0 and 1.
  + It is defined as and .

Where

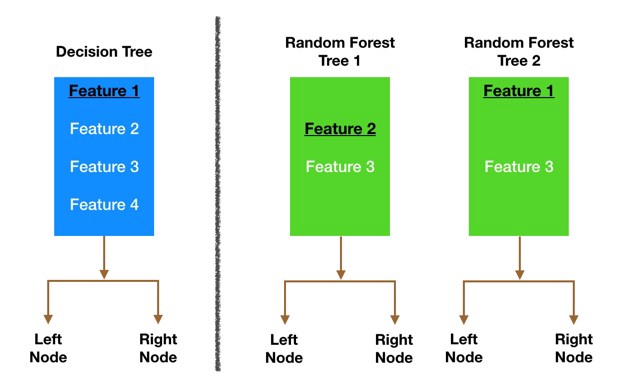
1. θis the model’s *parameter vector*, containing the bias term θ0 and the feature weights θ1 to θn.
2. θTis the transpose of θ(a row vector instead of a column vector).
3. xis the instance’s *feature vector*, containing x0 to xn, with x0 always equal to 1.
4. θT ・ xis the dot product of θTand x.
5. hθis the hypothesis function, using the model parameters θ.



* Second Model – Decision Tree
  + Decision Treesare versatile Machine Learning algorithms that can perform

both classification and regression tasks, and even multioutput tasks.

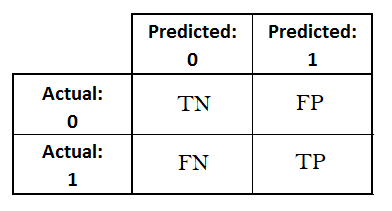
* + They are very powerful algorithms, capable of fitting complex datasets.
  + Decision Trees are also the fundamental components of Random Forests, which are among the most powerful Machine Learning available today.
  + A Decision Tree can also estimate the probability that an instance belongs to a particular class k: first its traverses the tree to find the leaf node for this instance, and then it returns the ratio of training instances of class k in this node.
  + Scikit-Learn uses Classification and Regression Tree (CART) Algorithm to train the decision tree.
  + The idea is quite simple: the algorithm first splits the training set in two subsets using a single feature k and a threshold tk.
* Third Model – Random Forest
  + Random Forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.
  + It needs features that have at least some predictive power.
  + The trees of the forest and more importantly their predictions need to be uncorrelated or at least have low correlations with each other.
  + Bagging (Bootstrap Aggregation) – Decision Trees are very sensitive to the data they are trained on – small changes to the training set can result in significantly different trees structures. Random Forest takes advantages of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees. This process is known as bagging.



* + Feature Randomness - In a normal decision tree, when it is time to split a node, it considers every possible feature and pick the one that produces the most separation between the observations in the left node vs. those in the right node. In contrast, each tree in a random forest can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.
  + The Random Forest creation pseudocode:
    1. Randomly select “**K**” features from total “**m**” features where **k << m**
    2. Among the “**K**” features, calculate the node “**d**” using the best split point
    3. Split the node into **child nodes** using the **best split**
    4. Repeat the **1 to 3** steps until “**l**” number of nodes has been reached
    5. Build forest by repeating steps **1 to 4** for “**n**” number times to create **“n” number of trees.**

Accuracy and Results:

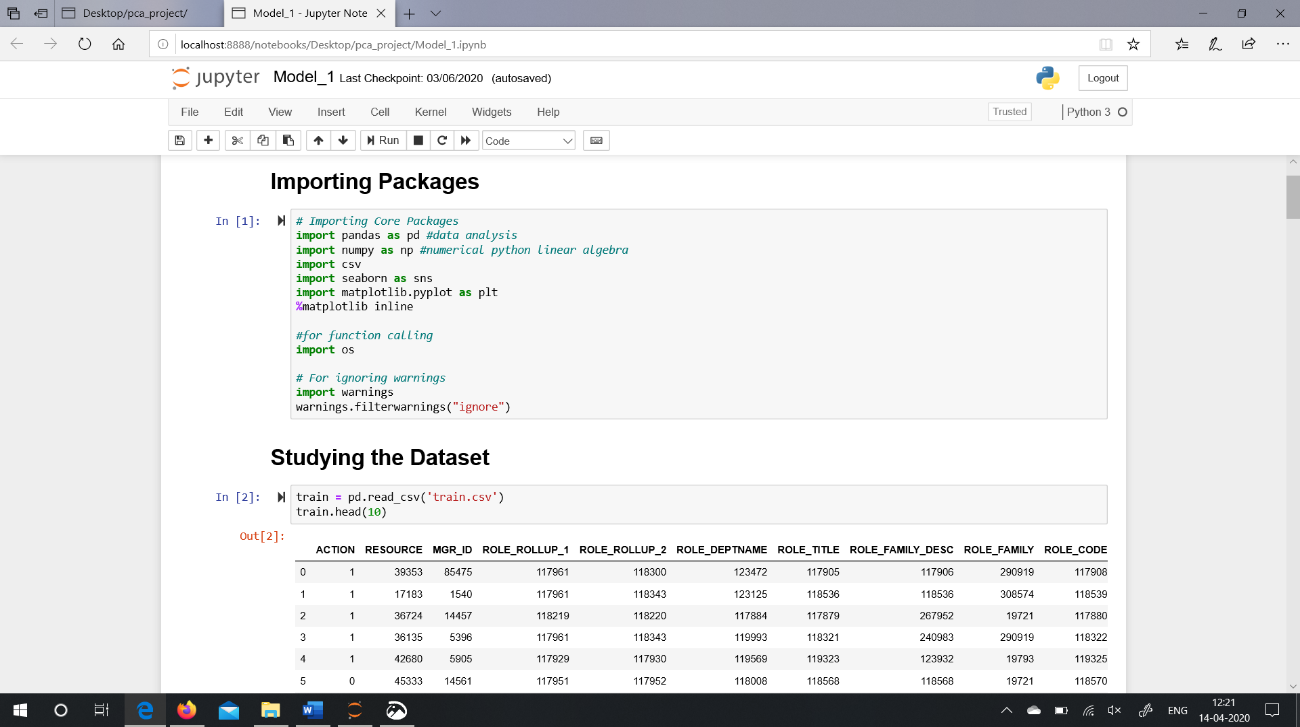
* Confusion Matrix is a table which is used to evaluate the performance of a classification model. It looks like the following.



* Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:
  + Accuracy =
  + For Binary Classification, accuracy can also be calculated in terms of positives and negatives as
  + Accuracy = where TP = true positives, TN = true negatives, FP= false positives, FN= false negatives.
* Precision is the fraction of relevant instance among the retrieved instances. In other words, it means that when the model predicts a positive value then what are the odds that the model has made a correct prediction. It is calculated as follows:
  + Precision =

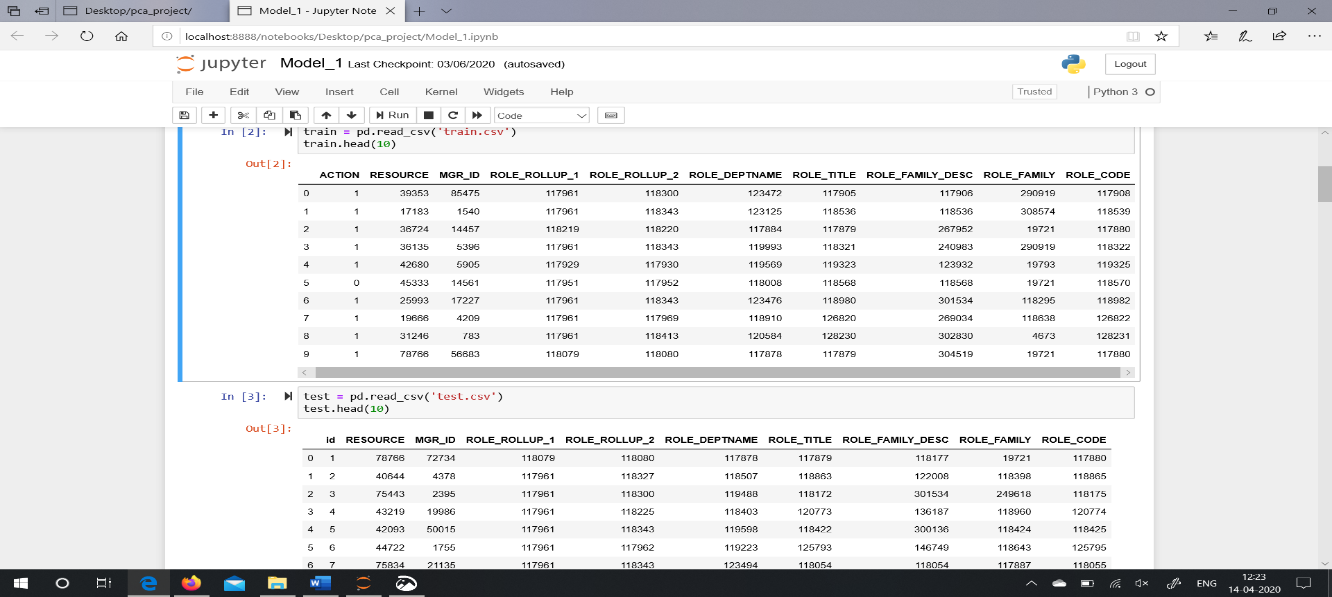
# Code Screenshots and Explanation

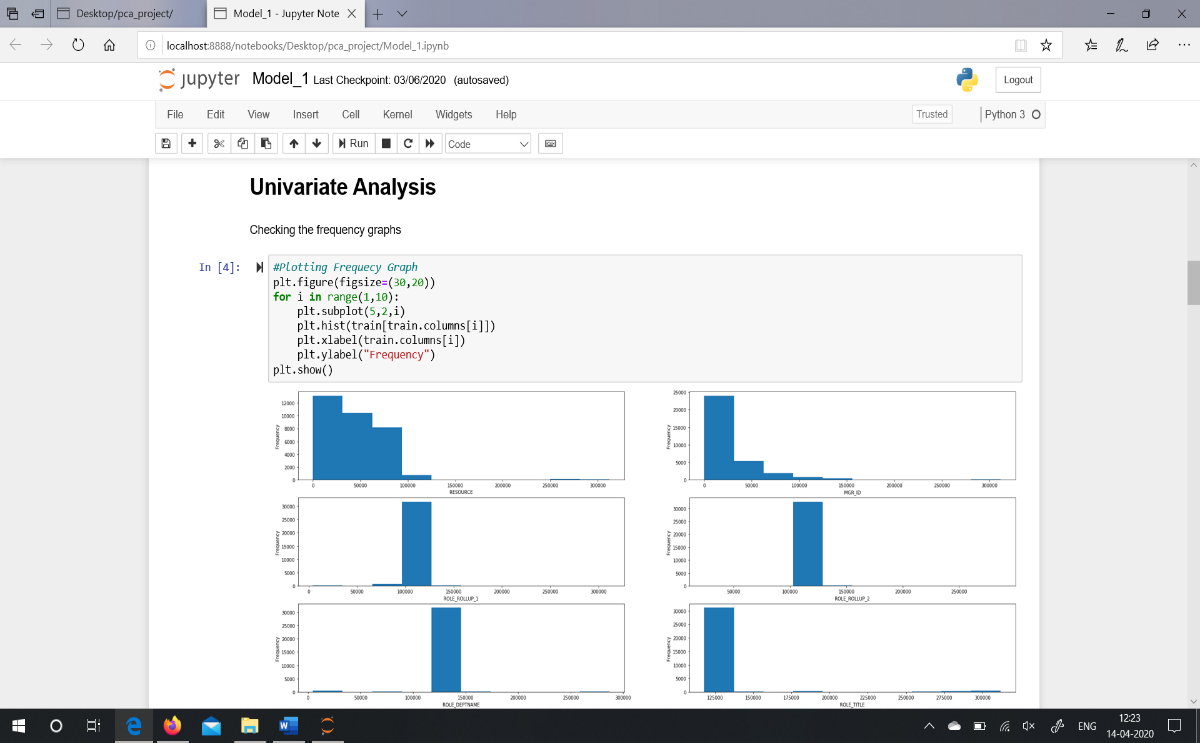
* Importing packages and datasets

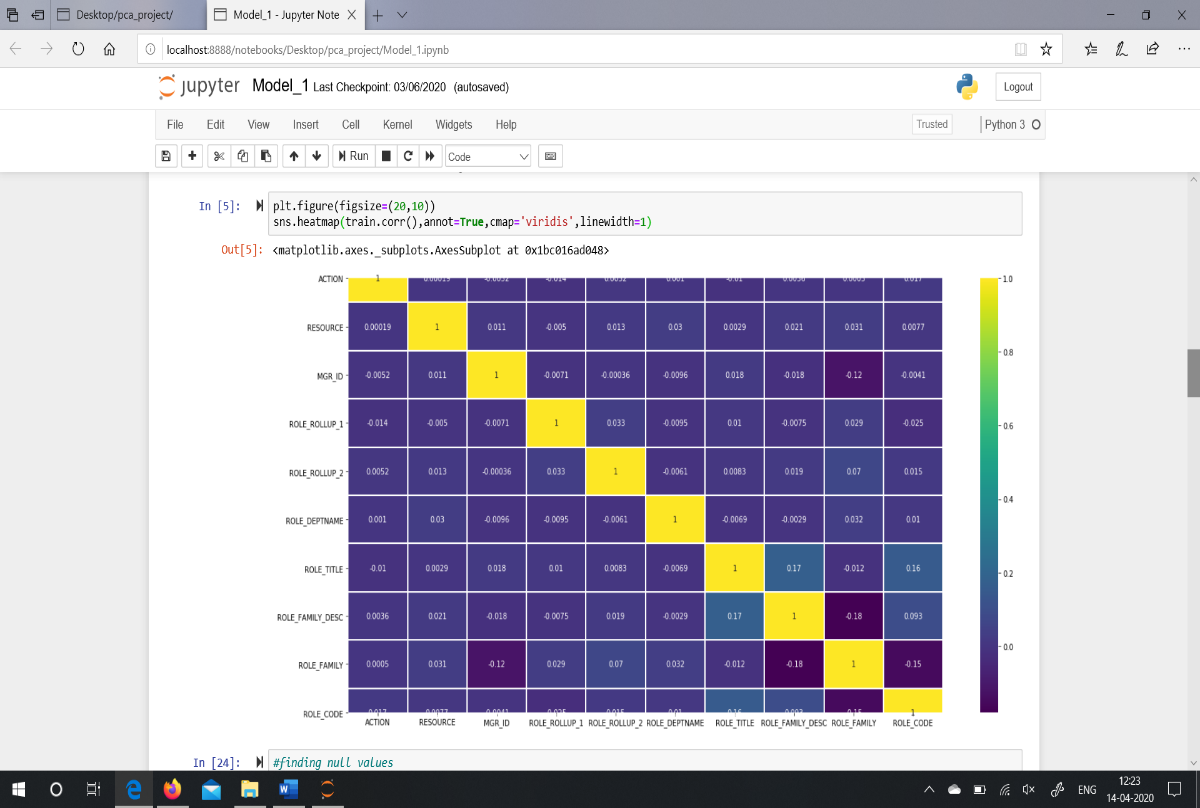


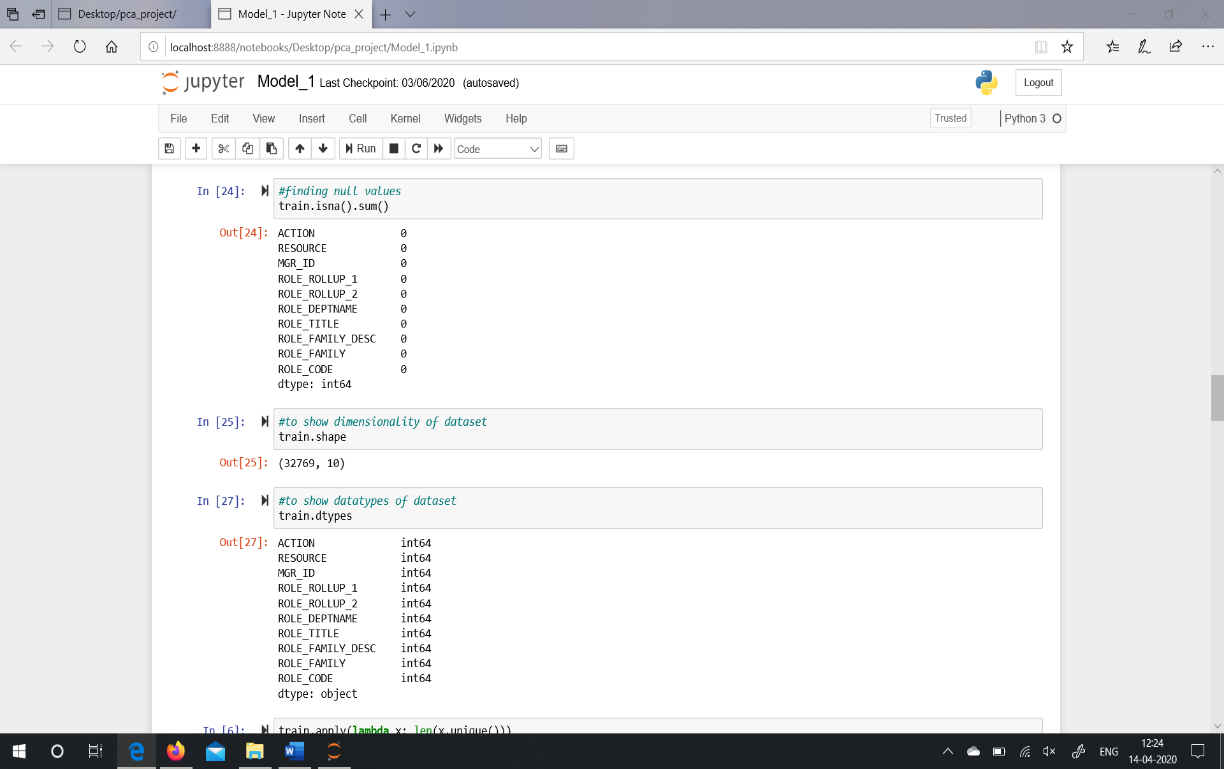
* Uni-variate Analysis Step

Step include checking of null values in dataset, data types of dataset, count of number of rows and columns, outlier treatment, frequencies of dataset, etc.



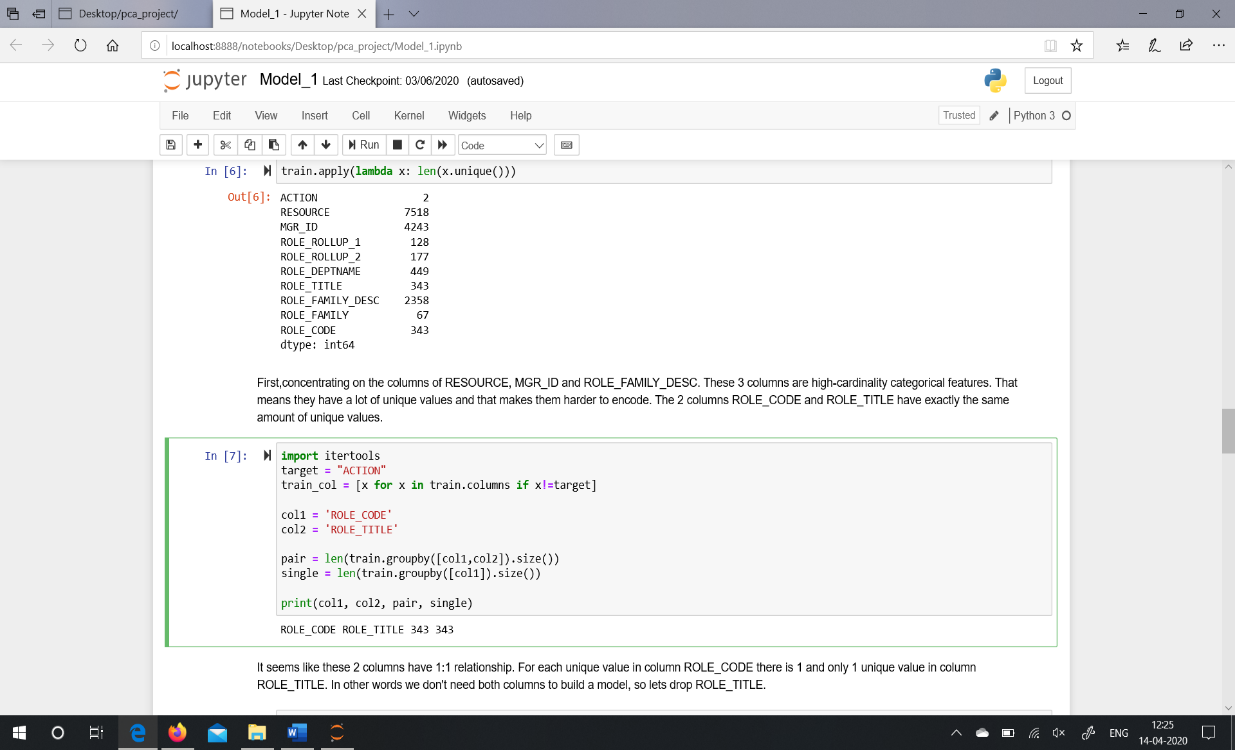




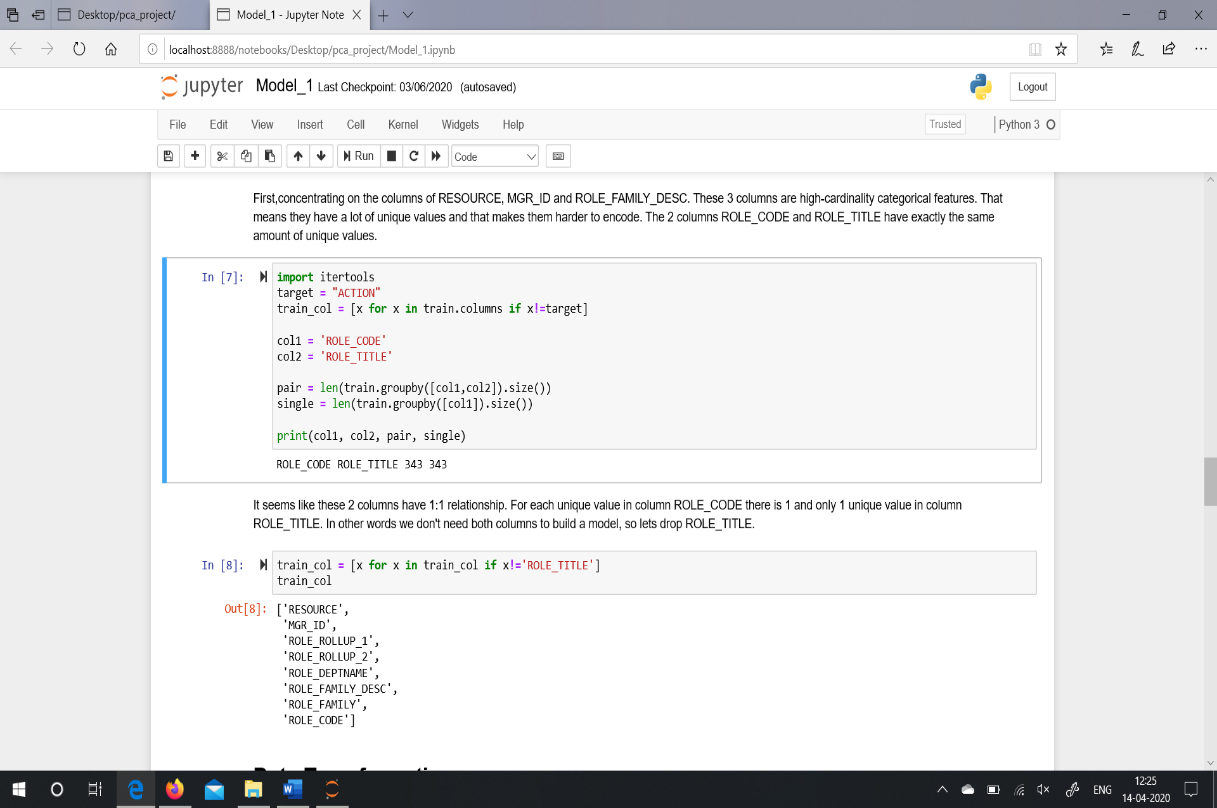


* Deleting the unnecessary column which are not required for modelling.

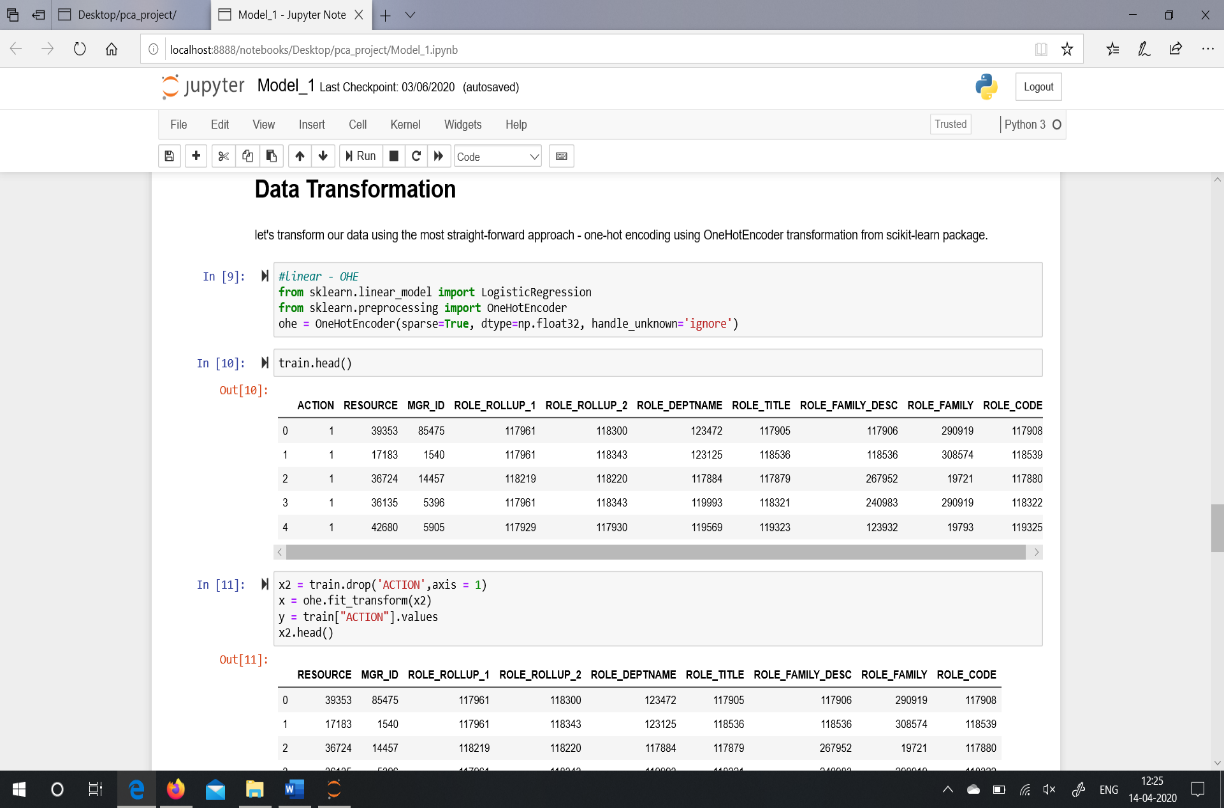
train\_col = [x for x in train\_col if x != ‘ROLE\_TITLE’]



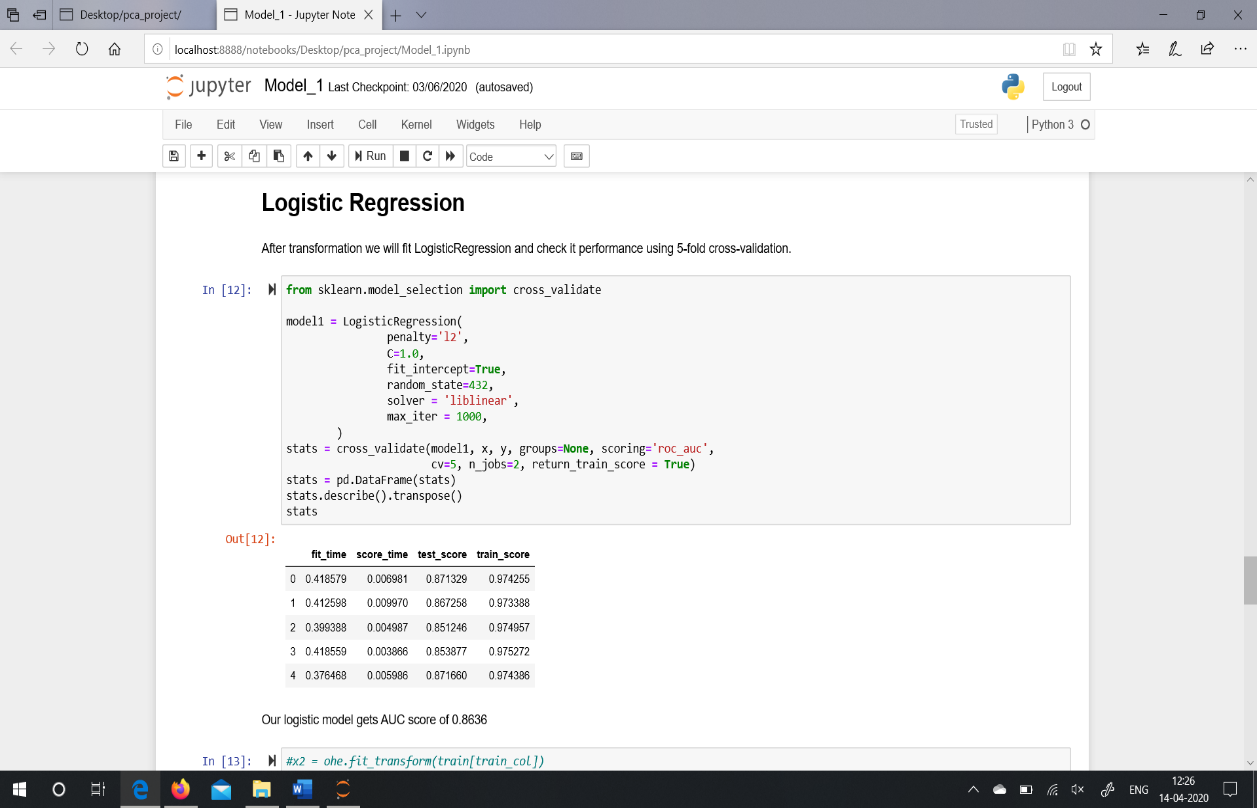
First, concentrating on the columns of RESOURCE, MGR\_ID and ROLE\_FAMILY\_DESC. These 3 columns are high-cardinality categorical features. That means they have a lot of unique values and that makes them harder to encode. The 2 columns ROLE\_CODE and ROLE\_TITLE have exactly the same amount of unique values and it seems like these 2 columns have 1:1 relationship. For each unique value in column ROLE\_CODE there is 1 and only 1 unique value in column ROLE\_TITLE. In other word there is no requirement of both columns to build a model, so drop ROLE\_TITLE.



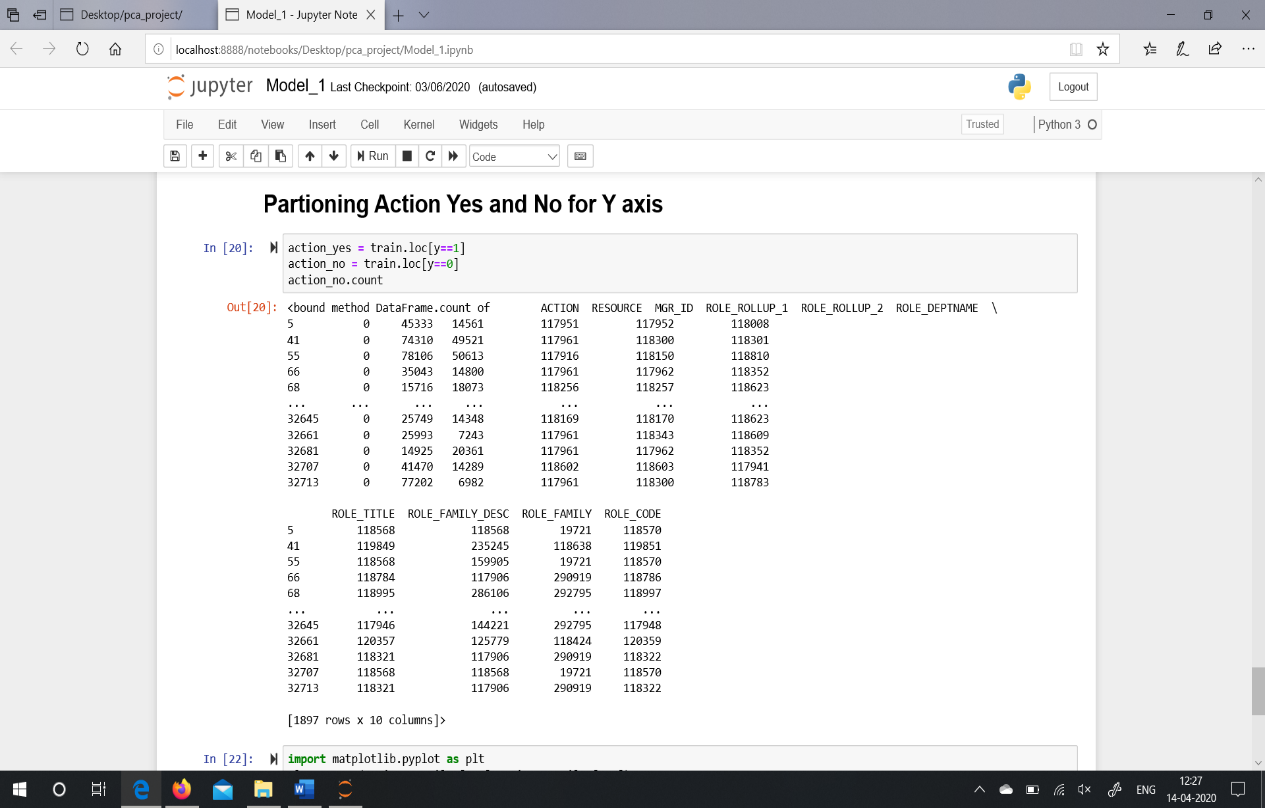
* Transforming dataset using One Hot Encoder approach.

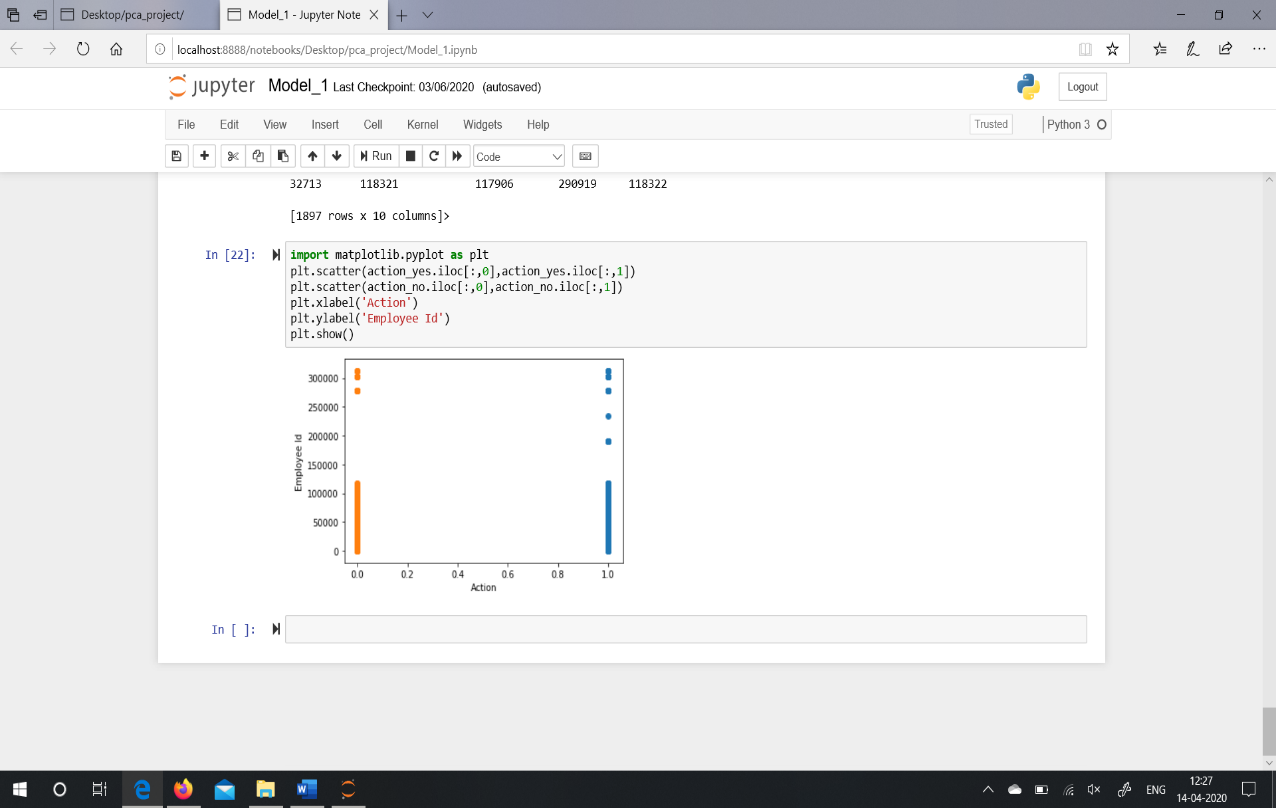


* Fitting Logistic Regression Algorithm



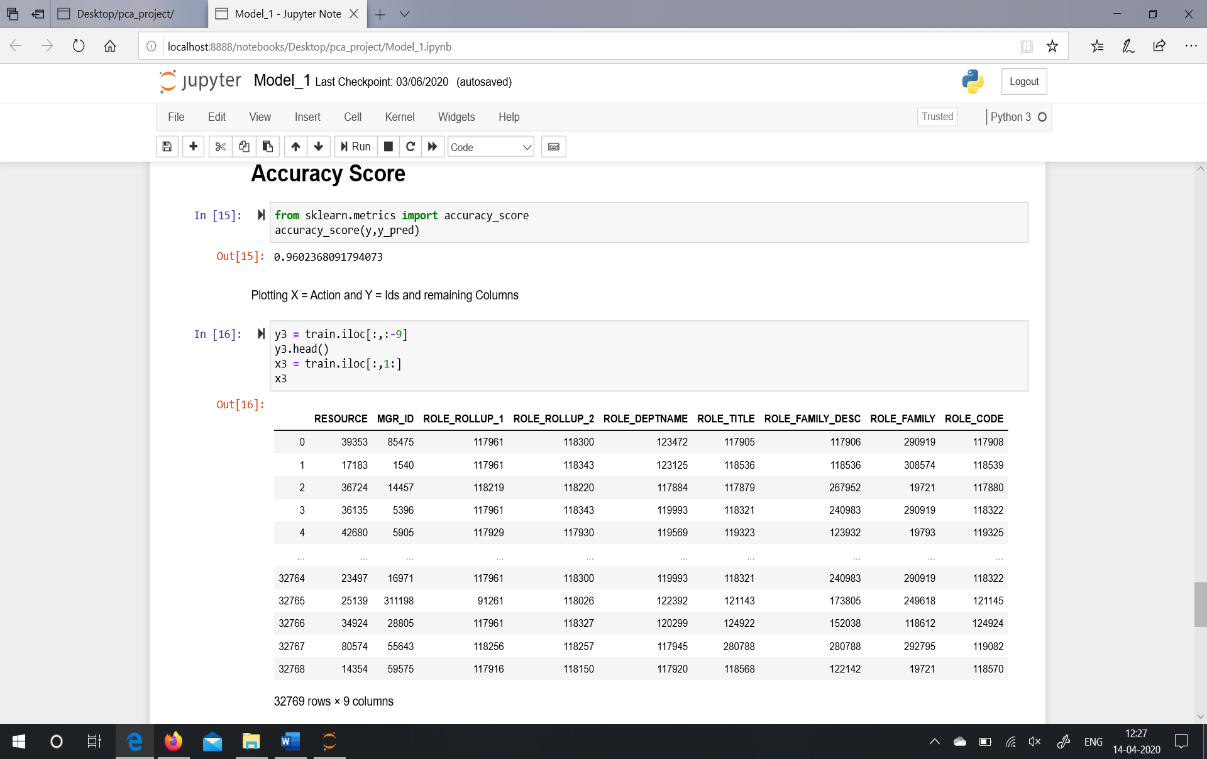
* Logistic Regression Algorithm Graph





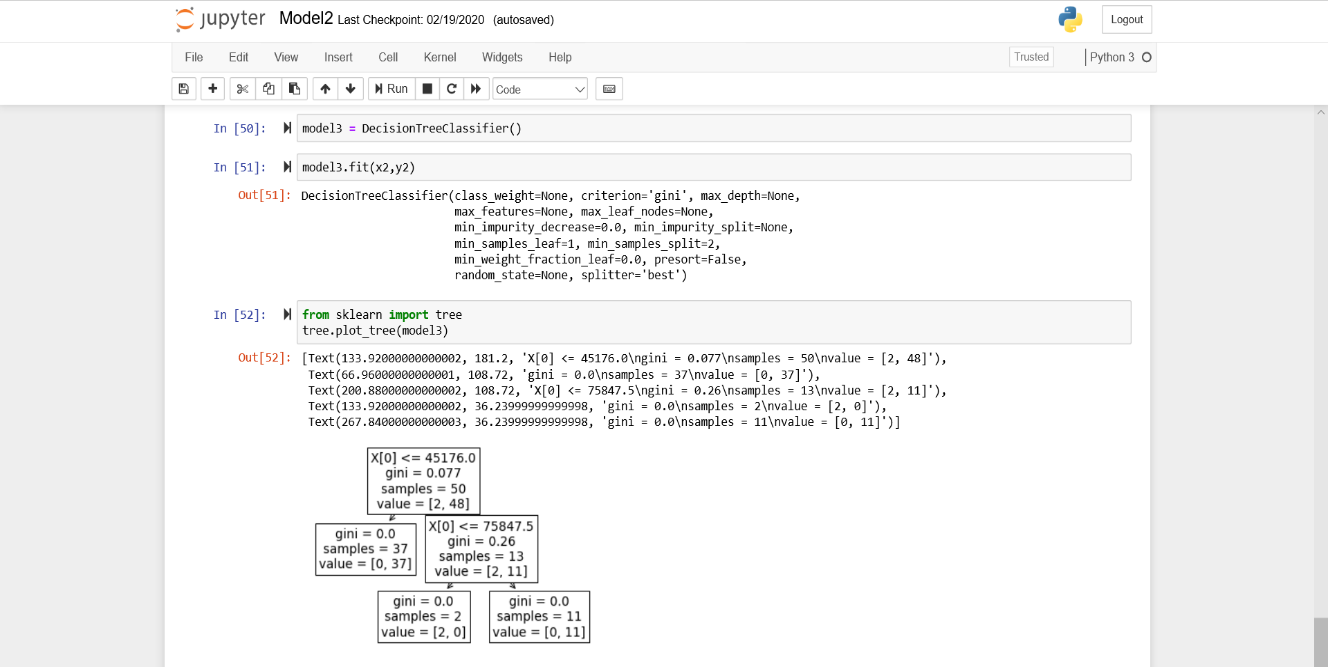
The graphs show the Action either “0” or “1” on X-Axis i.e., whether the resources were accessed or not accessed. Out of 32,769 employees 30,872 employees were given necessary access and remaining 1,897 employees were not given resources.

* Accuracy Score of Logistic Regression



Logistic Regression model shows 96% accuracy.

* Fitting Decision Tree Algorithm



This graph shows tree of only 50 datasets.

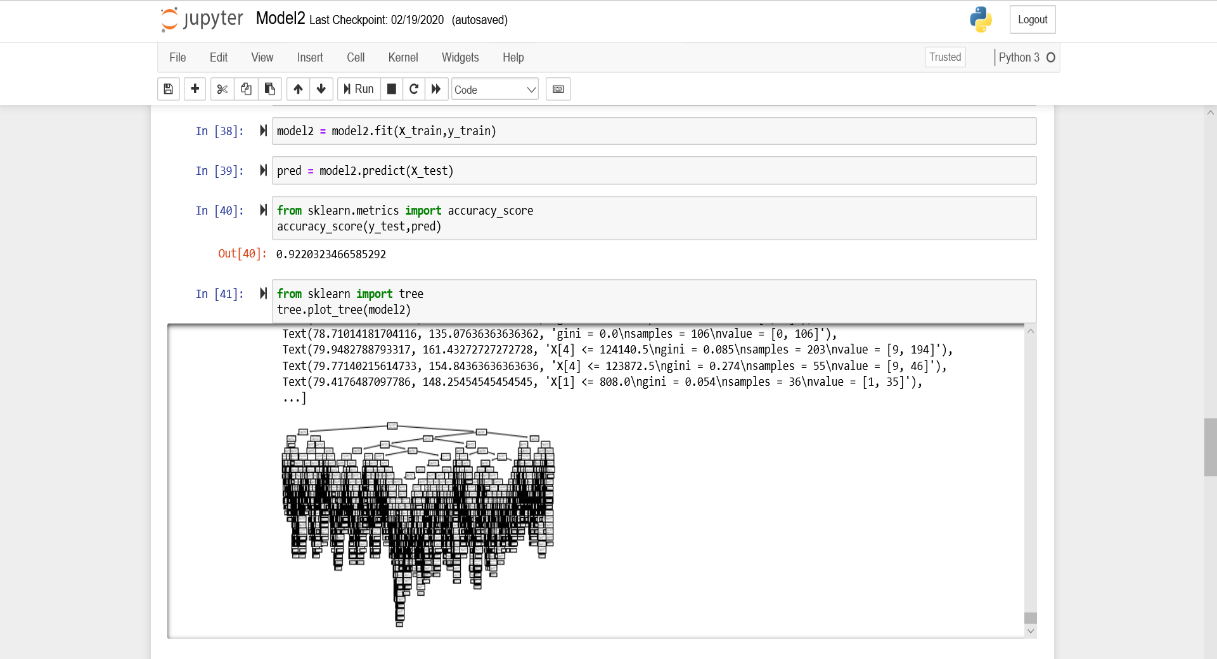
Criterion = gini: The Gini Impurity of the node. The average weighted Gini Impurity decreases as we move down the tree.

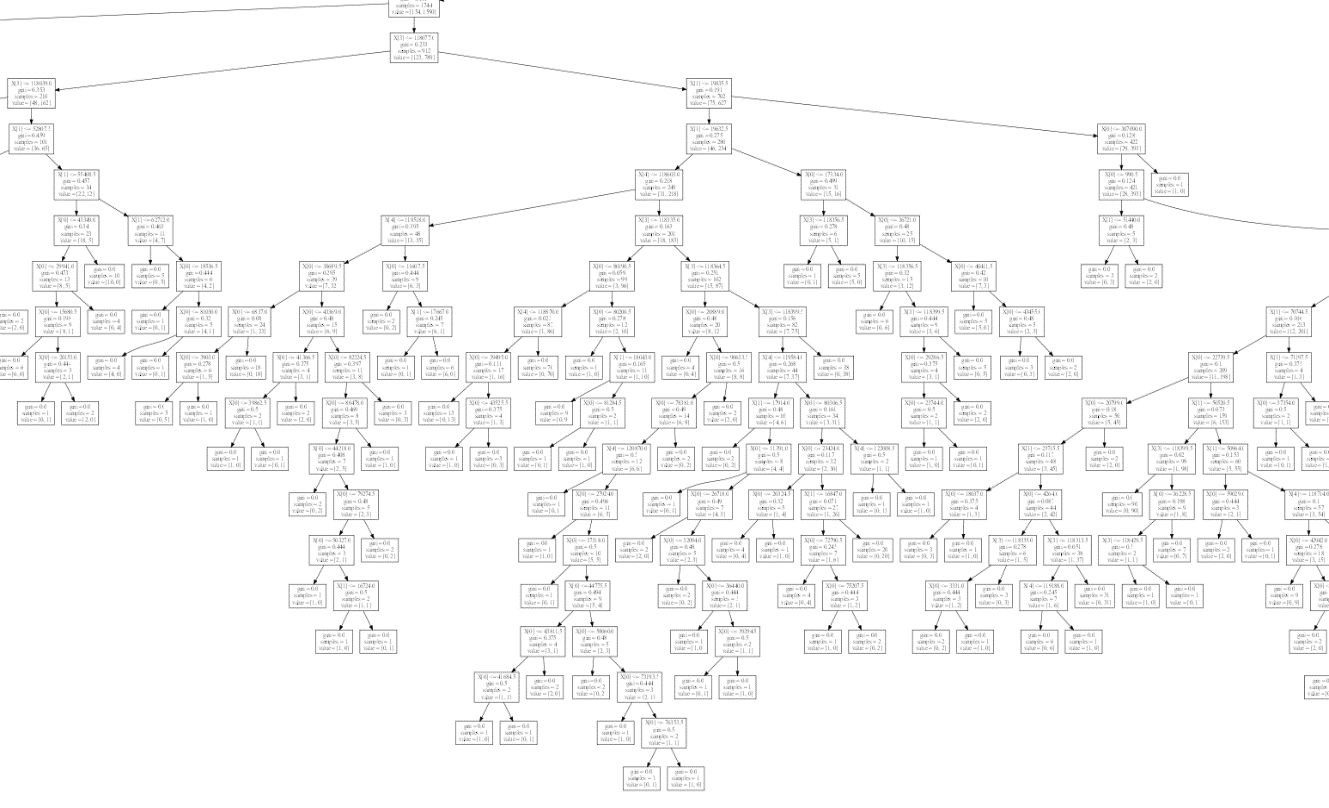
Samples: The number of observations in the node.

Value: The number of samples in each class. For example, the top node has 2 samples in class 2 and 48 samples in class 1.

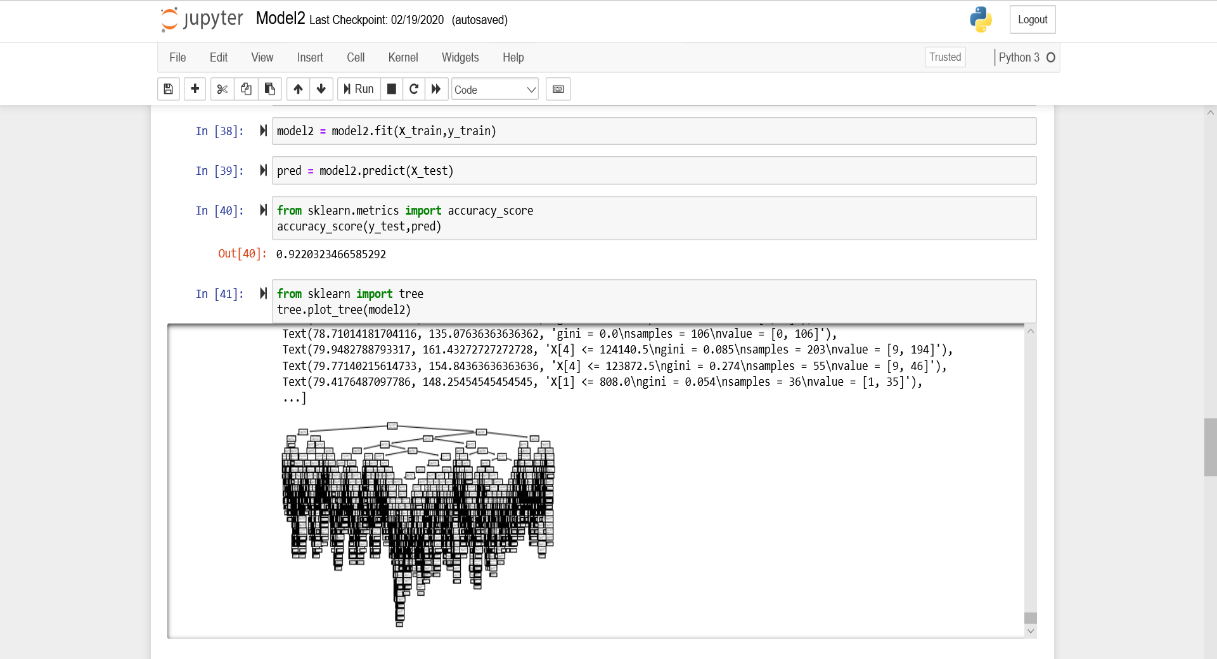
Class: The majority classification for points in the node. In the case of leaf nodes, this is the prediction for all samples in the node.

* Decision Tree Algorithm Graph



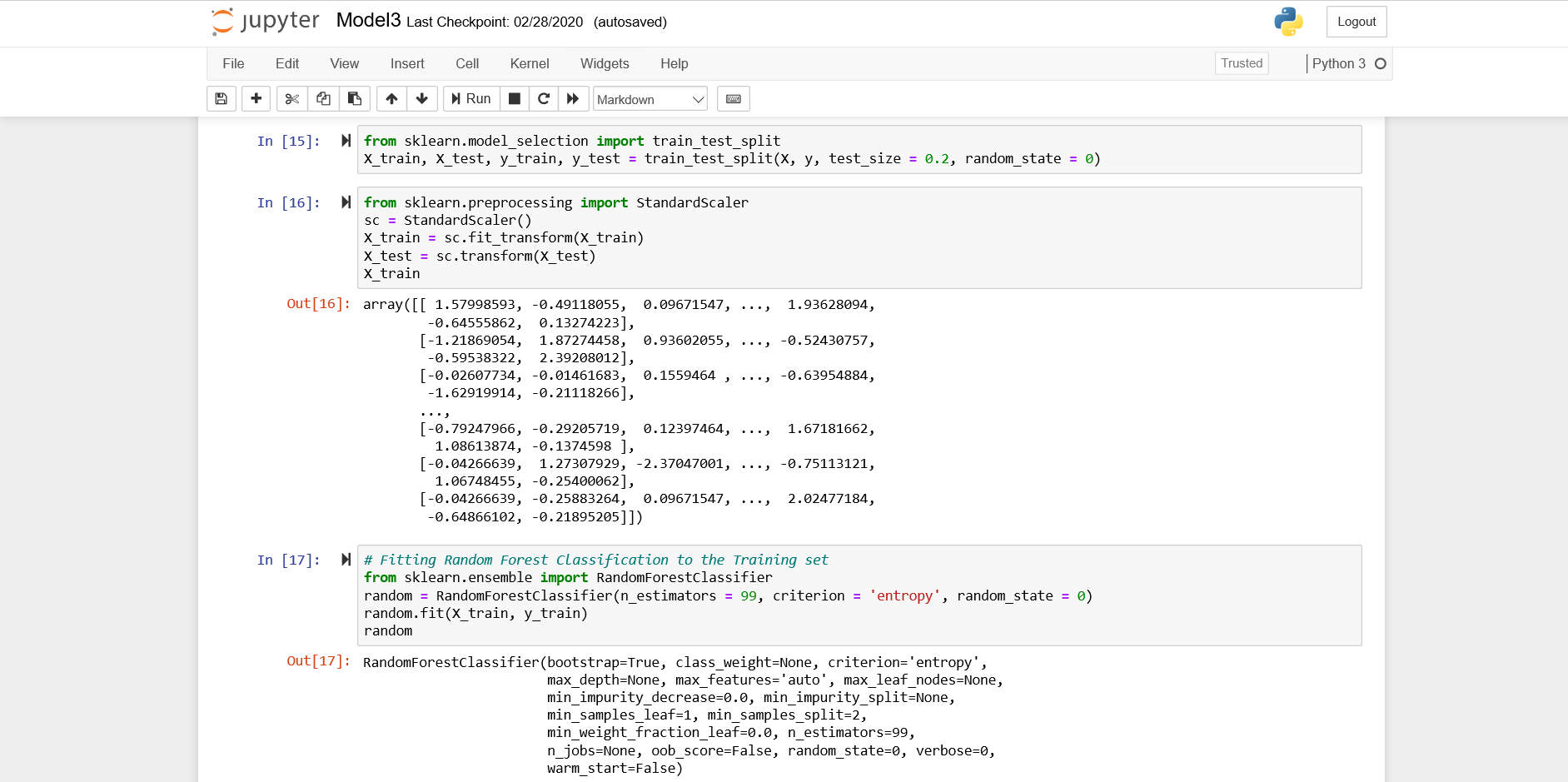


* Accuracy Score of Decision Tree Algorithm

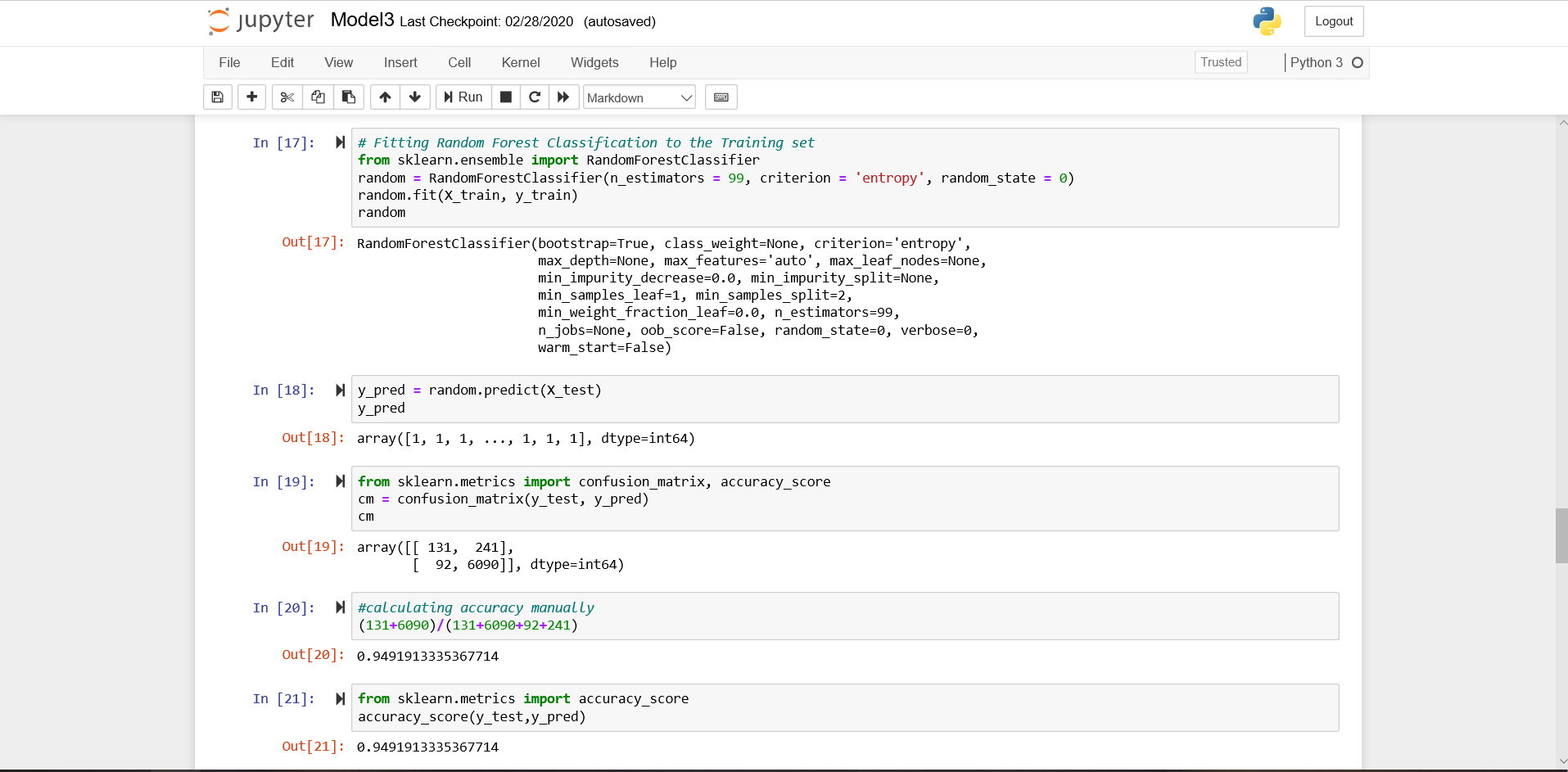


Decision Tree Algorithm model shows 92% accuracy.

* Splitting and transforming dataset using Standard Scalar approach.

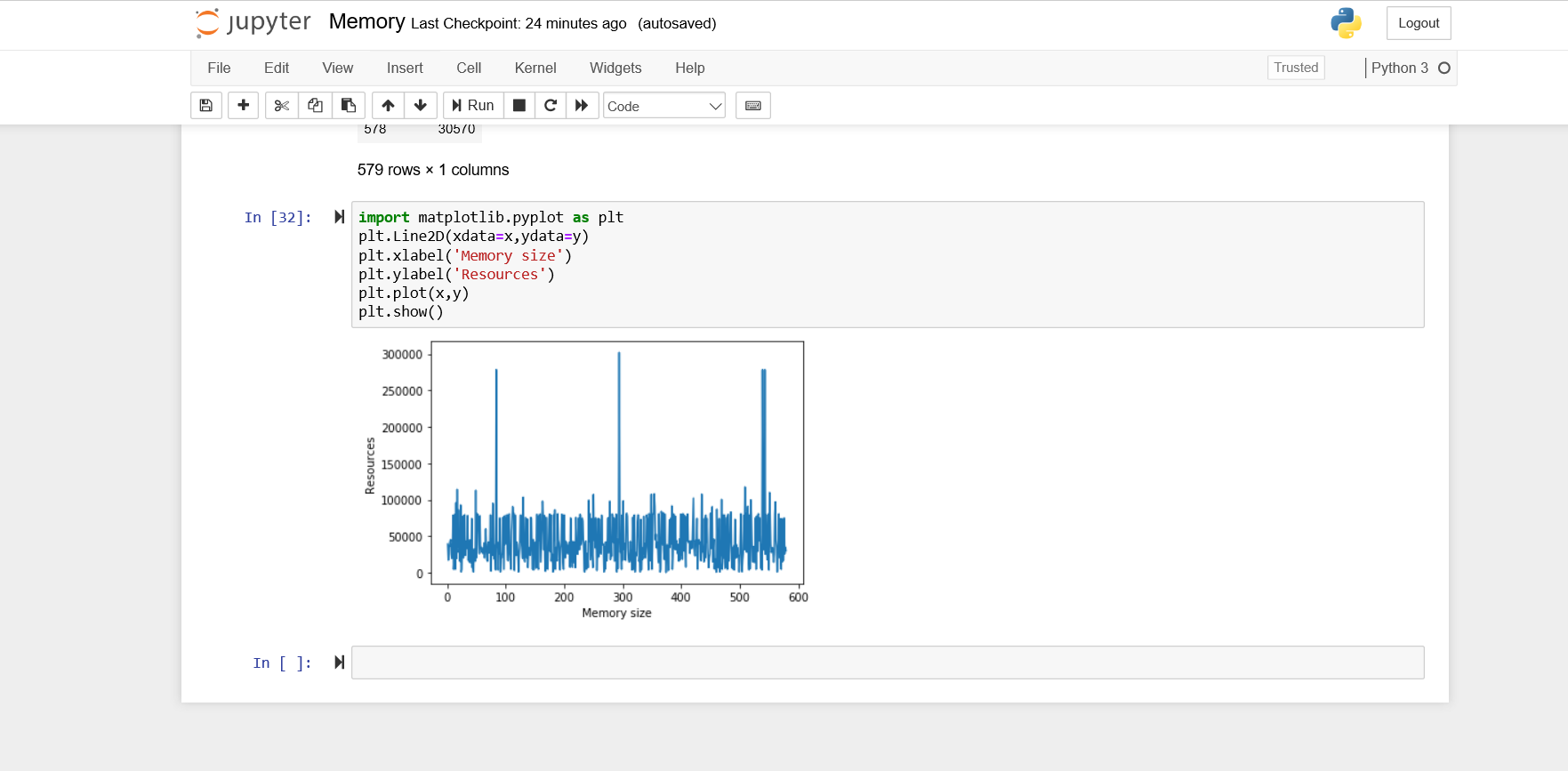


* Fitting and Accuracy Score Random Forest Algorithm



Random Forest Algorithm model shows 94% accuracy.

* Displaying Memory Size



To generate the memory size of the resources, using algorithm, size of memory is calculated by the combination of resources and the resources which are there in the various department. Resources lying at higher order in hierarchy will have more access compared to the lower level. In order to fulfil all the access higher resources will need more memory comparatively to lower resources.

# Testing

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. Each test type addresses a specific testing requirement.

## Blackbox Testing:

* Blackbox testing is testing the functionality of an application without knowing the details of its implementation including internal program structure, data structures, etc.
* Test cases for Blackbox testing are created based on the requirement specifications. Therefore, it is also called as specification-based testing.
* When applied to Machine Learning models, Blackbox testing would mean testing Machine Learning models without knowing the internal details such as features of the Machine Learning model, the algorithm used to create the model, etc.
* Model Performance
  + Testing model performance is about testing the models with the test data/new data sets and comparing the model performance in terms of parameters such as accuracy/recall etc., to that of pre-determined accuracy with the model already built and moved into production. This is the most trivial of different techniques which could be used for Blackbox testing.
* Metamorphic Testing
  + In metamorphic testing, one or more properties are identified that represent the metamorphic relationship between input-output pairs. For example, hypothetically speaking, an ML model is built that predict the employee access needs required based on different variables like Manager Id, Resource Id, Employee Id, Action and Target, etc.

* Dual Coding
  + With dual coding technique, the idea is to build different models based on different algorithms and comparing the prediction from each of these models given a particular input data set.
  + Let's day, a classification model is built with different algorithms such as random forest, logistic regression, decision tree. All of them demonstrate a comparative accuracy of 94% or so with random forest showing the accuracy of 97%. This results in the selection of random forest.

## Test Cases

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test cases | Pre-Conditions | Step to be Executed | Expected Result | Actual Result | Pass/Fail |
| Necessary Packages installed? | Packages for ml models and performing data analysis are required. | Install using command on your respective coding ground. | Success | Success | Pass |
| Check whether historical dataset is in same directory. | csv file should be in same directory. | If not in same directory, transfer it into it and then run the command pd.read\_csv | File should be read and displayed. | Appropriate message displayed. | Pass |
| Testing of Uni-variate analysis. | Checking all the necessary steps of univariate analysis. | Checking null value, data types of dataset, count of entries in dataset, number of rows and columns in dataset, etc | Null value checked, datatypes checked, entries count checked, rows and columns checked. | Success and displayed result. | Pass |
| Check transformed data. | Checking the transformed data and eliminating columns which are not required. | Using itertools, fit\_transform, standard scalar, etc | Columns eliminated and ready for training. | Success | Pass |
| Test if the models are fitted properly to the dataset. | Checking how many percent of data is taken from train set and test set and adding necessary parameters. | Importing packages from Scikit.learn of logistic regression, decision tree and random forest algorithm. | Return success and cross validation score. | Success | Pass |
| Test if the accuracy score is correct. | Generating and checking accuracy score of each algorithm is correct and valid. | Generating accuracy score by importing packages called as accuracy\_score or by constructing confusion\_matrix and then finding accuracy. | Every algorithm should get the score in between 94-97%. | Success | Pass |
| Test if every algorithm generates result graphs | Generating right and valid graphs for each algorithm. | Importing packages for graphs like matplotlib and seaborn, etc. | Scatter plot and tree graphs are shown | Success | Pass |

# Conclusion

* Our main contribution is focused in providing data analysis and model building for the dataset.
* Our project can be useful to develop a technology where it can predict an employee's access needs, such that manual access transactions (grants and revokes) are minimized as the employee's attributes change over time.
* Out of three machine learning models like logistic regression, decision tree and random forest algorithm, Logistic Regression shows the best accuracy of 96% and the rest Decision Tree shows 92% and Random Forest shows 94% accuracy.
* Our model is scalable and the datasets can be modified easily for different datasets accordingly.
* Our project model can be a part of building new software technology application used in companies to keep a track on employee’s resources access and providing adequate amount of memory to the employees depending on the employee’s role.
* Our model can also help in providing the information of unused resources access and auto-reset the whole computer access and provide to the new employees. This will help auto-access models to minimize the human involvement required to grant or revoke employee access.

References

* <https://www.kaggle.com/c/amazon-employee-access-challenge>
* <https://www.kaggle.com/c/amazon-employee-access-challenge/data>
* <https://docs.python.org/3.6/>
* <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>
* <https://towardsdatascience.com/preprocessing-with-sklearn-a-complete-and-comprehensive-guide-670cb98fcfb9>
* <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
* <https://www.dezyre.com/project/hackerday-project/project-title/amazon-employee-access-needs#sub-about-hackerday>
* Intro to Machine Learning by Ethem Alpaydin
* Hand on Machine Learning, O’Reilly