**REPORT ON CAPSTONE PROJECTS**

**Abstract**

**Project 1**

**CREDIT CARD CUSTOMER SEGMENTATION**

A Bank wants to develop a customer segmentation to define marketing strategy. The goal of this analysis report is to discover the Customer Segmentation of a bank, by looking through their behaviour/profile while using Credit Card. Hopefully, we can get a clear segmentation of the customer, so we can deploy effective marketing campaign or sales promotion to the targeted costumer.

**Project 2**

**EMPLOYEE ATTRITION PREDICTION**

Employee turn-over (also known as "employee churn") is a costly problem for companies. The true cost of replacing an employee can often be quite large.

Understanding why and when employees are most likely to leave can lead to actions to improve employee retention as well as possibly planning new hiring in advance. This project would fall under what is commonly known as "HR Analytics", "People Analytics". Given that we have data on former employees, this is a standard supervised classification problem where the label is a binary variable, 0 (active employee), 1 (former employee). In this study, our target variable Y is the probability of an employee leaving the company.

**Tools/Skill Used**

1. Python Programming
2. Python Libraries and Packages Numpy, Pandas, Matplotlib, Seaborn
3. Data Visualisation
4. Exploratory Data Analysis
5. Machine Learning

**Project -1 Credit card customer segmentation:**

**INTRODUCATION**

Segmentation in marketing is a technique used to divide customers or other entities into groups based on attributes such as behaviour or demographics. It is useful to identify segments of customers who may respond in a similar way to specific marketing techniques such as email subject lines or display advertisements. As it gives businesses the ability to tailor marketing messages and timing to generate better response rates and provide improved consumer

experiences.    

**OBJECTIVE**

This analysis aims that a bank wants to identify segments of customers who may respond in a similar way to specific behaviour. I have to develop a customer segmentation to define marketing strategy. The sample Dataset summarizes the usage behaviour of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioural variables. This study also aims to see different Machine Learning Methods on the same data set.

**MEDOTHLOGY**

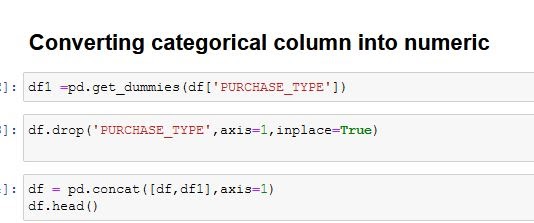
* Data Description
* EDA and Visualisation of the Data
* Build a Clustering Model

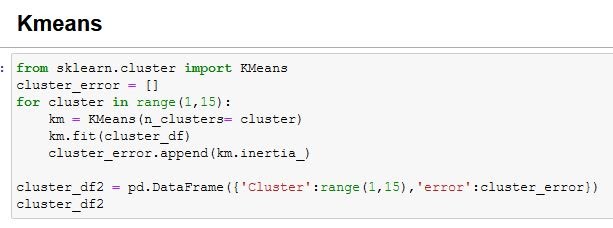
**DATA DESCRIPTION**

* Name of the data: CREDIT-CARD-SEGMENTATION.csv
* Features- 18
* Rows- 8950

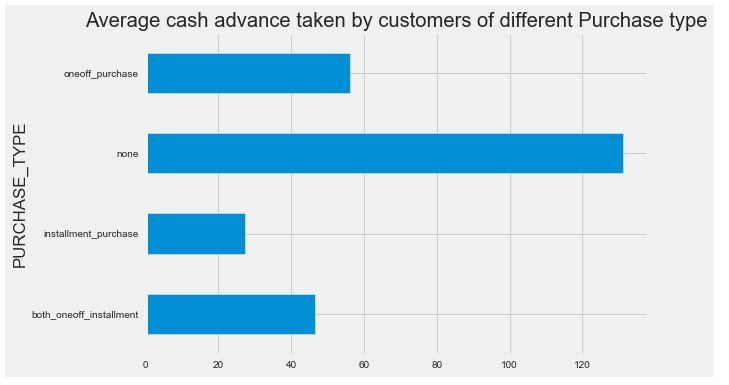
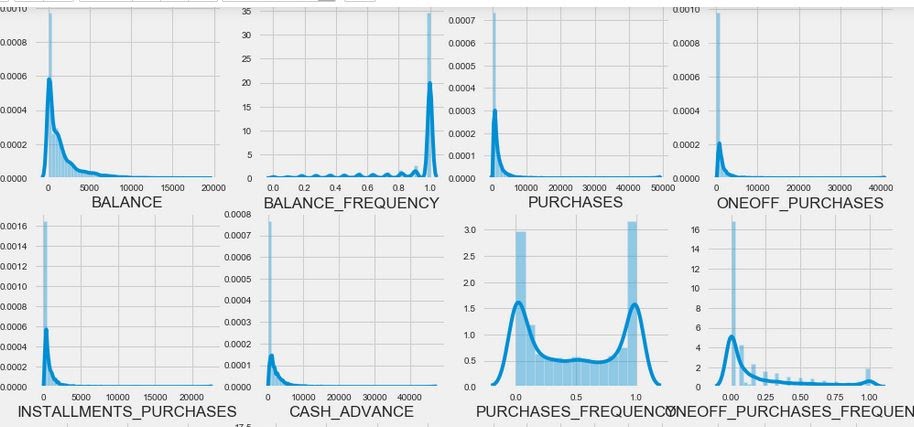
**VISUALISATION OF THE DATA**

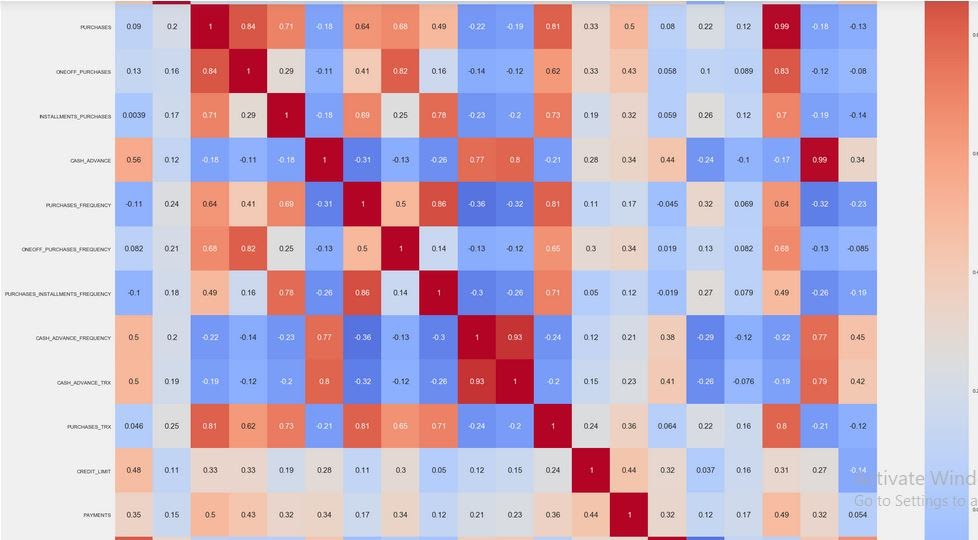
**Code Snippets:**

****

****

**Visualization Snippets:**

****

****

**MODELLING**

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 than those in other groups. In simple words, the aim is

to segregate groups with similar traits and assign them into clusters.

1. **KMeans -** Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

**CONCLUSION/RESULT**

As a result, we can say, Kmeans with 6 clusters is the best fit to this dataset, giving the best separation of data. Data set gives six types of customers in which the customers which take maximum advance\_cash and have maximum credit\_score & doing no purchase transaction are maximum. This group is about 23% of the total customer base.

**FUTURE SCOPE**

As the company generates more data on its employees (on New Joiners and recent Leavers) the algorithm can be re-trained using the additional data and theoretically generate more accurate predictions to identify high-risk employees of leaving based on the probabilistic label assigned to each feature variable (i.e. employee) by the algorithm.

**Project 2 - Employee Attrition Prediction**

**INTRODUCATION**

Employee turn-over (also known as "employee churn") is a costly problem for companies. The true cost of replacing an employee can often be quite large.

Understanding why and when employees are most likely to leave can lead to actions to improve employee retention as well as possibly planning new hiring in advance.

**OBJECTIVE**

The aim of this study to identify the below problem and make a prediction model ::

* What is the likelihood of an active employee leaving the company?
* What are the key indicators of an employee leaving the company?
* What policies or strategies can be adopted based on the results to improve employee retention?

**MEDOTHLOGY**

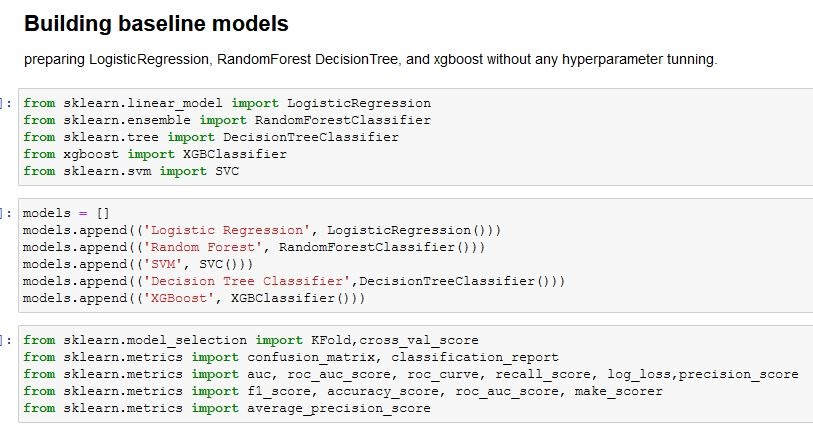
* Data Description
* EDA and Visualisation of the Data
* Build a Classification Model
* Choose the correct model

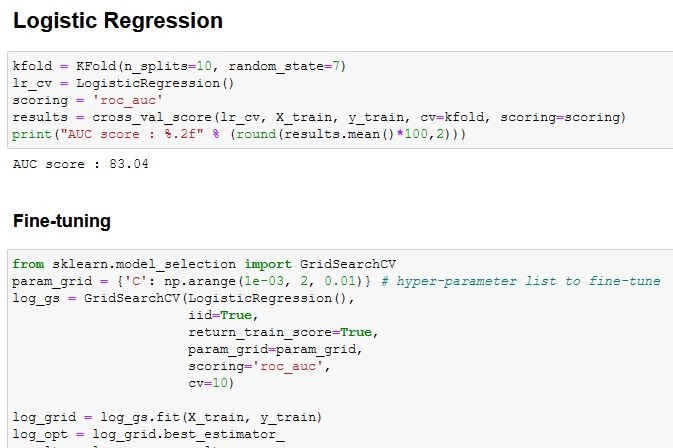
**DATA DESCRIPTION**

* Name of the data: Employee-Attrition.csv
* Number of data points: 1470
* Number of features: 35
* Target attribute: Attrition
* Range of target attribute: Range of target attribute: (Yes, No), Yes=16%, No=84%

**VISUALISATION OF THE DATA**

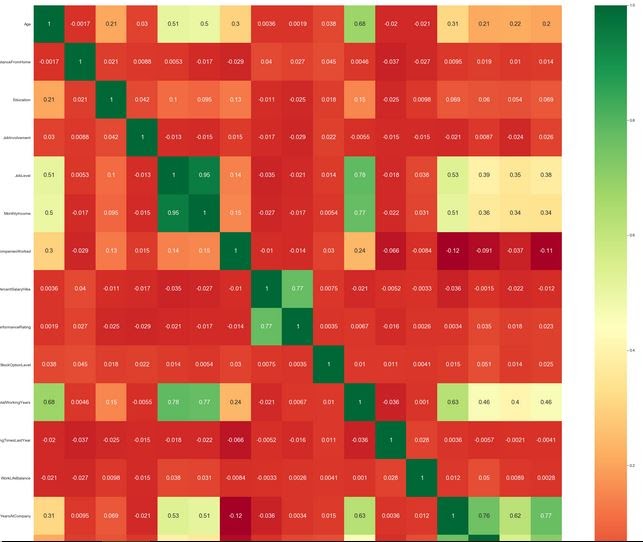
**Code Snippets:**

****

****

**Visualization Snippets:**

****

****

**Modeling:**

**1. Logistic Regression:** The logistic model is used to model the probability of a  certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.  This can be extended to model several classes of events such as determining whether  an image contains a cat, dog, lion, etc.

**2. Random Forest:** Random forests or random decision forests are an ensemble  learning method for classification, regression and other tasks that operate by  constructing a multitude of decision trees at training time and outputting the class that  is the mode of the classes or mean prediction of the individual trees.

**3. Support Vector Machine:** Support vector machines (SVM) are a set of supervised  learning methods used for classification, regression and outlier’s detection. Support  vector machines are supervised learning models with associated learning algorithms  that analyze data used for classification and regression analysis.

**4. Decision Tree:** A decision tree is a decision support tool that uses a tree-like model  of decisions and their possible consequences, including chance event outcomes,  resource costs, and utility. It is one way to display an algorithm that only contains  conditional control statements.

**EDA Insights:**

1. The workers with low JobLevel, MonthlyIncome, YearAtCompany, and TotalWorkingYears are more likely to quit their jobs.
2. **Gender** : The male employees are likely to quit their job but the percentage of both seems to be equally likely.
3. **Marital Status:** Single people are more likely to leave than others.
4. **BusinessTravel :** The workers who travel a lot are more likely to quit then other employees.
5. **Department :** The workers in Research & Development are more likely to stay then the workers in other department.
6. **EducationField :** The workers with Human Resources and Technical Degree are more likely to quit then employees from other fields of education.
7. **JobRole :** The workers in Laboratory Technician, Sales Representative, and Human Resources are more likely to quit the workers in other positions.
8. **OverTime** : The workers who work more hours are likely to quit then others.

**Conclusion/ Results**

As a result we can say, Logistic Regression is the best fit to this dataset, gives 73%  accuracy and 0.72 ROC -AUC score with fine tuning .  Random Forest Models is also working very well on this Dataset. It gives  87% of accuracy and 0.70 ROC\_ AUC score which is quite good. I choose ROC-AUC as evaluation metrics  because data is highly imbalanced so Logistic Regression is the best model for this data set.

**FUTURE SCOPE**

 With feature importance we get to know that in future when these models are picked up, then we don’t need to run the model on all the variables only those top 10 parameters are enough to predict the cancer as we have seen that with these 10 parameters also the accuracy remains the same. In future, the models can be upgraded with some better techniques in terms of getting higher and better accuracy, f1-score, precision, recall etc.

