**Business Analytics Portfolio Project:- Dive Deep on the absenteeism module of a company**

**Introduction:-**

In this portfolio project we will take a real time dataset that constitutes about the absenteeism of a company and apply Data cleaning, Ml modelling techniques and visualize the data to better understand the use case and propose business solutions to minimize and resolve the same.

**Analytics Techniques used**

|  |  |
| --- | --- |
| **Technique** | **Description** |
| Software Integration in SQL | To store the dataset in SQL and retrieve the base in python (for developing model) and Tableau (for visualizations) |
| Data Cleaning | To dive deep on the data and cleaning practices such as resolving missing values, feature dropping, feature select and feature engineering |
| Applied ML modelling | Perform a train test split for the cleaned data, and estimate the accuracy of model prediction |
| Data Visualization | To compute Key features and present interactive visualizations to leadership for simplified understanding. |

**Tools Used:-**

|  |  |
| --- | --- |
| **Tool** | **Purpose** |
| Excel | Features observation, high level analysis and pivots |
| Python | For Exploratory Data analysis, Data cleaning, ML model selection, training, testing and storage |
| SQL | To store the CSV file obtained from python operations and for future integration and use |
| Tableau | Visualization the key parameters and obtain inferences of the case study |

**Case Study Overview from a business perspective:-**

The dataset provides us inputs on absenteeism during the work hours of a given company. Absence from work results in temporary incapacity during normal working hours to execute a regular work activity. Our goal is to apply business analytics techniques and derive information/proposals that will help in preserving /increasing the productivity rates of a company

**Workaround:-**

**The following analytical approaches will be applied to derive solutions**

* Data cleaning
* Applied ML modelling
* Visualizations

**Step 1:- > Data cleaning**

* The below details contain a step by step approach to perform cleaning practices using the **pandas and** **numpy** python libraries and get the dataset ready for ML modelling
* **Dropping** the ID column, as this parameter contains unique entries of the employees will have a minimal role in the diving deep and understanding the use case agenda
* The target **prediction feature** for this model will be the “reason for absence”, the column contains numbers (as **categorical features**) flagged for each reason. For example: 1 –> Sick leave, 2-> Personal Emergency, 3 -> Have an unscheduled commitment etc.
* We will apply **dummy variables** to split the categorical column reason, and analyze the data. The table will be split into 27 numbers, which each number contain a unique reason of absence.
* The next step is **grouping the dummy variables.** The dataset as of current contains many parameters, our goal will be optimize the model by concatenating the 27 parameters into groups, this will help avoid to simplify data interpretation and improve the model training.

The reason are group as per the below table

|  |  |  |  |
| --- | --- | --- | --- |
| **Group 1** | **Group 2** | **Group 3** | **Group 4** |
| Certain infectious and parasitic diseases  Neoplasms  Blood diseases  Endocrine and metabolic diseases  Mental disorders  Nervous Disease  Eye Disease  Ear Problem  Circulatory system  Respiratory system  Skin and subcutaneous  Genitourinary system | Pregnancy, childbirth  Perinatal period  Chromosomal abnormalities | Abnormal clinical and laboratory findings  Injury , poisonings  Mortality and Morbidity  Health status, health services | Patient follow-up  Medical consultation  Blood donation  Laboratory examination  Unjustified absence  Physiotherapy  Dental consultation |

* The next step is to work with the date column and change the data type to date using the date time function and extract the week and month using weekday formula and month formula.
* We will then analyze the education column, this consists of 4 values (1 to 4). The number holds the degree ranking of the employee.

|  |  |
| --- | --- |
| **Number** | **Degree** |
| 1 | High School |
| 2 | Graduate |
| 3 | Post Graduate |
| 4 | Masters or Doctorate |

1 -> High school

2 ->Graduate

3 -> Post Graduate

4 -> Master or Doctorate

* On further analysis of the counts for each group using the **value\_counts** function we can interpret that the high school group is more prevailing than the rest. To optimize the model we will recreate this as a classifier which depicts 1 for High School and 0 for the rest of the degree.

**SQL Integration**

* In this section we will store the cleaned Absenteeism in an SQL database for optimal storage. We will then use ‘**pymsql’** to connect with the jupyter notebook and write python code for ML modelling.
* The first step is to upload the derived clean csv file into the SQL interface. We will then create a database with all the mentioned columns for storage. Let us name the DB as “predicted outputs”
* The ‘pymsql’ package is imported and a connection is setup mapped with the path of the file. We will then create a cursor to write SQL code in a python interface.
* Open the “predicted outputs” using the Select function. The connection is successfully setup we can now write SQL queries for analysis and start building the ML model.

**ML Modelling**

* In this phase, we will input the cleaned data set and apply ML modelling techniques to predict the accuracy of the model.
* Creating the targets for the logistics regressions, the targets will be people of 2 classes, where class 1 is the people who have been absent more than median of class 1 and 2
* On calculation the median using the **median** function we can comprehend the results as, people unavailable for more than 3 hours is considered as absent.
* The next step is to select inputs for the regression, our predicting parameter will be target and the rest will fall as training parameters. We will apply a **train\_test\_split** in 80 to 20 ratio and then predict the accuracy.
* The model will be tested with the testing data and the accuracy of the model performance will be gauged.
* The trained model will be saved as a **pickle** file for further use. The output derived from this will be utilized for visualization and retrieve inferences to propose our solutions.

**Tableau Visualizations**

In this section we will import the csv file to the tableau interface and performance data visualizations to retrieve inferences.

The following relations will be evaluated

* Age vs Probability
* Reasons vs Probability
* Transportation vs Probability

**End inferences**

* The higher the transport expense and children count, the greater chances of leave availability
* The frequency of health issues and medical checkups is proportionate to the rate of absence
* The probability increases as the age increases.