Submission Date: 15 Oct 2023 11.59 PM

Weightage: 30%

Indian industries are giving direct employment to millions of workers. Acquiring a new talent is always a key concern of the organizations, especially when they don't have adequate experience in identifying the right talent and appointing him/her to the appropriate role. In many cases, it is difficult to find an exact match for the job specified. If an offer is denied, then the human resource (HR) department has to repeat the entire recruitment process resulting in additional effort.

That's why many of the organizations depend on the third party specialist service providers who have wide experience in hiring processes across different industries. "HRMagic" is one of the known names in this space having a huge client base spread across different sectors like IT, manufacturing, finance etc. over the years they have realized the pain organizations face with respect to the hiring process. At same time they have collected a huge dataset of the recruitment drives they have conducted for the various organizations. While working on different recruitment drives, they have observed that 30% of the candidates who accept the job offer, do not join the company. This leads to huge loss of revenue and time as the companies need to initiate the recruitment process again to fill in the workforce demand. While looking for the solutions for this problem, the top management has come across the use of predictive analytics techniques applied by foreign talent management firms. They found those techniques quite interesting and have asked the team to explore more on the same.

"HRMagic " team has decided to proceed with a prototype using which they want to find out if a model can be built to predict the likelihood of a candidate joining the company. If the likelihood is high, then only the company will go ahead and offer the jobs to the candidates. They have already extracted the portion of the dataset from their huge pile of data which contains several attributes about candidates along with a column that indicates if the candidate finally joined the company or not. The detailed description of the other attributes is provided along with the dataset file.

For that purpose they have reached out to several groups who can help the "HRmagic" team to build the required model. Yours is one of them. After careful evaluation of the models received from several groups, the management will take a call on whether to go ahead and invest more into this predictive modeling strategy or not.

As a part of this project, you will be making use of datasets provided and help building the model. The main objective of this project is to give you real life experience while doing data acquisition, data integration, data cleaning and data transformation before attempting any of the analytical activity on the data.

The various tasks that you will be doing as a part of this exercise will be as follows:

ME
RG
EF
OR

t / extract data

The first step in building a model is to collect or extract data on the dependent variable and independent variables from different datasets provided. Data collection / extraction is a time consuming and expensive process.

- As a team you have to merge the candidate data provided based on the common factors.
- simultaneously you need to search what are the attributes those are taken into consideration while carrying out recruitment activity.
- The research will provide you more insight on determining the useful attributes present in the dataset or give pointers to the variables that need to be derived. make a list of such potential attributes.
- The outcome of this step will be a merged data set containing the 200 candidates data.
- document all your efforts appropriately in the jupyter notebooks with description and code.
- the weightage for this task will be 5 marks

(B) pre-process data

Before the model is built, it is essential to ensure the quality of the data for issues such as reliability, completeness, usefulness, accuracy, missing data and outliers.

Data imputation techniques may be used to deal with missing data. use of descriptive stat and visualization may be used to identify the existence of outliers and variability in the dataset. Many new variables can be derived and also used in model building. categorical data has to be pre-processed using dummy VARIABLES, before it is used to model building.

- you have to cleanse your datasets to remove all such daunting issues.
- narrate all the issues which you encounter during this exercise clearly with appropriate explanation and code.
- The weightage for this task will be 7 marks.

(C) perform descriptive analytics on data

It is always good to perform descriptive analytics before moving to building a predictive analytics model. it will help to understand the variability in the model. It also helps in identifying the relationship between the variables present in the dataset.

- use the exploratory data analysis technique on the dataset in order to find out the interesting insights that are hidden within the data captured.
- describe all eda steps those are done with the observations obtained out of it with the help of python code in jupyter notebook.
- the weightage for this task will be 7 marks.

(D) feature engineering

e Engineering is an important step to develop and improve performance of Machine Learning models. these techniques can also be used to identify the variables that impacts the outcome of the model. its basically Process of identifying and extracting the useful features from the available data. the Primary goal is to derive a set of features that best represent the insights hidden in the data, with a simpler model that generalizes well to future (unknown) observations.

- In this step you have to use knowledge of feature selection methods to identify the variables that have greater impact on the outcome.
- An elaborate description of the impact observed is expected as an outcome of this step.
- The weightage for this task will be 6 marks.

(E) model building and diagnostics

In this stage first data is divided into train and test data. The subset can also be created using random / stratified sampling procedure. This is an important step to measure the performance of a model using a dataset not used in model building. It is also essential to check for any overfitting of the model. the model is built using a training dataset to estimate the model parameters. the method of classifications can be utilized for the same.

- you have to prepare a logistic regression model to predict the probability of a candidate joining the company. assume "OfferRejected" as positive cases and "joined" as negative cases.
- find the significant features from the above model and build another logistic regression model with only the significant variables.
- compare the performance of both models using various model attributes and recommend a model that can be used by "HRmagic"
- the weightage for this task will be 5 marks

Notes:

- This is a take-home PRoject to be carried out by a group OF LEARNERS.
- as per the need, the demos / vivas can be arranged further on.
- wherever required you can make appropriate assumptions but make sure that you have spelt them appropriately in the submitted documents.
- This is a programming exercise - requiring the approach of appropriate model BUILDING.
- You may consult / discuss with other learners peripheral aspects such as the environment but not on solving the specific problems in terms of design or implementation.
- You have to write the appropriate Python code in Jupyter notebook to support you answers and submit with following nomenclature

  - FE_project1_<group_ID>.ipynb
- In case of any further queries, if those are generic ones, learners are encouraged to use discussion forums, otherwise they can reach out to me at ppawar@wilp.bits-pilani.ac.in.
- Manage your efforts properly as there is no scope to shift the deadlines announced above.

1) Logistics Regression
2) Logistics Regression Python tutorial
3) Logistics Regression explained