# AIML Feature ENGINEERING Projet 1

Submission Date: ~~03 June 2020 11.59 PM~~

Weightage: 30%

Indian industries are giving direct employment to the 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 at appropriate role. in many cases, it is difficult to find exact match for the job specified. if offer is denied, then the human resource (HR) department has to repeat the entire recruitment process resulting into additional effort.

that’s why many of the organizations depends on the third party specialist service providers who has wide expereince in hiring process acroos different industries. “HRMagic” is one of the known name in this space having a huge client base spread acroos different sectors like IT, manufacturing, fianance 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 recruitements drives they have conducted for the various organizations. while working on different recritment drives, they have observed that 30% of the candidates who accept the jobs offer, do not join the company. this leads to huge loss of revenue and time as the companies needs to initiate the recruitement process again to fill in the workforce demand. while looking for the solutions for this problem, the top management has came acroos use of predictive analytics techniques applied by foreign talent management firms. they found those techniques quite interesting and has 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 liklihood of a candidate joining the company. if the liklihood is high, then only company will go ahead and offer the jobs to the candidates. they already have extracted out the portion of 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 carfull evaluation of the models received from several groups, the management will take a call whether to go ahead and invest more into this predictive modelling 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 acquistion, data integration, data cleaning and data transforamtion 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:

1. collect / 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 time consuming and expensive process.

* You as team of three or four students you have merge the candidate data provided based on the common factors.
* simultanesuoyly you need to make 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 varibles that needs to be dervied. make list of such potential attribtues.
* outcome of this step will be merged data set containing the 200 candidates data.
* document all your efforts appropriately in the jupyter notebooks with description and code.
* the weigtage for this task will be 5 marks

1. pre-process data

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

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

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

1. 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 weigtage for this task will be 7 marks.

1. feature engineering

Feature 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.
* ellaborate description of the impact observed is expected as outcome of this step.
* the weigtage for this task will be 6 marks.

1. model buildingand 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 important step to measeure the performance of model using dataset not used in model building. it is also essential to check for any overfitting of the model. then model is built using training dataset to estimate the model parameters. the method of classifications can be utilized for the same.

* you have to prepeare a logistic regression model to predict the probablity of a candiate 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 performace of both models using various model attributes and recommend a model that can be used by “HRmagic”
* the weigtage for this task will be 5 marks

Notes:

* This is a take-home PRoject to be carried out by group OF LEARNER.
* as per the need, the demos / vivas can be arranged further on.
* wherever required you can make appropriate assumption but make sure that you have spelt them appropriately in the submitted documents.
* This is 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 once, learners are encouraged to use discussion forums, otherwise they can reach out to me at [ppawar@wilp.bits-pilani.ac.in](mailto:ppawar@wilp.bits-pilani.ac.in).
* Manage your efforts properly as there is no scope to shift the deadlines announced above.

References:

1. [Logistics Regression](https://www.kdnuggets.com/2020/03/linear-logistic-regression-explained.html)
2. [Logistics Regression Python tutorial](https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8)
3. [Logistics Regression explained](https://realpython.com/logistic-regression-python/)