UDACITY

Udacity Machine Learning Engineer Nanodegree 2020

Capstone Proposal

Predicting Employee Attrition in the Dawn of Recession

Chetan S Rane

October 2020

Domain Background:

As the COVID-19 keeps unleashing its havoc, the world continues to get pushed into the crisis of the great economic recession, more and more companies start to cut down their underperforming employees. Companies firing hundreds and thousands of Employees is a typical headline today. Cutting down employees or reducing an employee salary is a tough decision to take. It needs to be taken with utmost care as imprecision in the identification of employees whose performance is attriting may lead to sabotaging of both employees' career and the company's reputation in the market.

Problem Statement:

The goal is to predict the employee attrition by the given data about his/her past history.
From the given data we have to select best features using statistical and visual analysis which help in better classification of a given record. Trying feature engineering can be very useful in strong feature creation which help in better classification of employee attrition.
The challenging part about dataset is we have a good variety of categorical features.
Also considering the number of feature and records we have, one more challenge here is of curse of dimensionality.

Acknowledgements:

We thank IBM for providing us with the dataset.

Dataset:

The dataset contains 1628 rows and 29 columns on which we will be training our data.

Data Fields:

- Id - an anonymous id given to an Employee
- Age - Age of an Employee
- Attrition - Did the Employee leave the company, 0-No, 1-Yes
- BusinessTravel - Travelling frequency of an Employee
- Department - Work Department
- DistanceFromHome - Distance of office from home
- EducationField - Field of Education
- EmployeeNumber - Number of Employees in the division of a given Employee
- EnvironmentSatisfaction - Work Environment Satisfaction
- Gender - Gender of Employee
- MartialStatus - Martial Status of an employee
- MonthlyIncome - Monthly Income of Employee in USD
- NumCompaniesWorked - Number of Companies in which Employee has worked before joining this Company

- OverTime - Does The person work overtime
- PercentSalaryHike - Average annual salary hike in percentages
- StockOptionLevel - Company stocks given to an Employee
- TotalWorkingYears - Total working experience of an employee
- TrainingTimesLastYear - No. of trainings an employee went through last year
- YearsAtCompany - Number of years worked at this company
- YearsInCurrentRole - Number of years in current role
- YearsSinceLastPromotion - Number of years since last promotion
- YearsWithCurrManager - Number of years with the current manager
- Education
- EnvironmentSatisfaction
- JobInvolvement
- JobSatisfaction
- PerformanceRating
- Behaviour
- CommunicationSkill
- StockOptionLevel

Please find the below Kaggle link to dataset. Metadata and file information can be found on the below link.

https://www.kaggle.com/c/summeranalytics2020/data

Solution Statement:

Data Pre-processing:
Check for missing values, outliers, data cleaning, check for class imbalance and it's impact, handling of categorical variables.

Model building:
Starting with basic classification model, will move to tree-based model and then ensemble models if performance found unsatisfactory.
As we don't know which algorithm and combination of hyperparameters will be best suited for dataset to give best generalized results, we have shortlisted below model.

1) Logistic Regression
2) Support Vector Machine
3) Gaussian Naïve Bayes
4) Decision Tree
5) Random Forest
6) Gradient Boosting
7) Bagging Classifier

To select the best hyperparameters we will use Random search cross validation technique as it is very fast compared to Grid search cross validation technique.

Project Design:

- **Data Pre-processing:**
  Check for missing values, outliers, data cleaning, check for class imbalance and its impact, handling of categorical variables.
- **Data Splitting:**
  Split the data into training set and validation set with and 80-20 split
- **Model Training and evaluation:**
  I will start with simple model architecture first before training and evaluating it. Then iterate this process trying different architectures and hyper-parameters to reach an accuracy score we are happy with.

Benchmark Model:

This dataset is labelled dataset which typically falls under supervised classification problem. For this kind of problem tree-based model will perform much better than rest of the models. Finally, it all depends on distribution of data with respect to target class. So, we will pick Random Forest model as the benchmark and try beat the benchmark with other models along with hyperparameter tunning.

Evaluation Metrics:

The evaluation metric for this competition is Confusion Matrix, Classification report and AUC score under ROC.

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

In our case we need to predict the attrition of employee which is our positive class. So our main goal will be to reduce False negatives i.e actual is attrition but predicted as no attrition.