Machine Learning Engineer Nanodegree

Capstone project

Arshad Khurshid
Oct 21, 2017

Using Supervised Learning to identify attrition in Company

Domain Background

Machine Learning Engineer Nanodegree course has emphasized on different type of Supervised and Unsupervised Learning. Supervised Learning deals with Regression and Classification. Classification is being implemeted across all the fields for different applications and are very prominant in Cyber Security and Medical diagnosis. With boom in data, Classification model has gained momentum to understand the cause of change in behaviour in consumer. This helps business to work on data and not to take decision just on intitution.

A company invest lot on employee to train them and make them ready for next generation business. Once you invest in skill enhancement of an employee you need to used it for benefit of business. Employee may be agitated even if they are being paid well as human have aspiration and if aspiration is fulfilled then they perform to their maximum capability. I was looking for dataset which has features that can be reason for employee to quit and Kaggle gave me one. Retaining an employee means retaining knowledge and they are the one who grooms the people working in one level down thus helping to increase and knowledge base for all.

Dataset Link: https://www.kaggle.com/ludobenistant/hr-analytics (https://www.kaggle.com/ludobenistant/hr-analytics)

Problem Statement

Managing people in a company is challenging job. When it comes to attrition HR has a tough job to identify the root cause of attrition. Most of the time it happens that diagnosis based on intitution is worng and hence the policy implemented has more damaging effect and situation worsen. What if HR knows in that particular employee is more probable of leaving and could take damage control action. What is they know why employee are leaving and features that is having more damaging effect on them. Since i need to emphasize on 2 groups Classification of employee into Left and Stayed and i treat it as Classification problem. The other is the reason why people are Leaving. For this i will drop the decision feature(left) and will group features in different nodes and find the features that has most impact.

Features and description

- satisfaction_level : Level of Satisfaction
- last_evaluation :Time since Last performance Evaluation
- number_project :Number of Project completed while at work
- average_montly_hours : Average monthly hours at workplace
- time_spend_company : Number of years spent in the company
- Work_accident : Whether the employee had a workplace accident
- left : Whether employee left the workplace or not
- promotion_last_5years: Wehter employee was promoted in last 5 years
- sales : Department they work for
- Salary : Relative level of Salary(high)

Model will be predicting if person is potential employee who will stay or leave the company. Satisfaaction level is important attribute and a person not satisfied will probably leave the organization. last_evaluation means that performing person is not getting enough feedback and hence enthusiasm to continue work is lost. number_project does not seem important attribute as project can be small or big. Some employee may have worked on big projects continued for long time whereas some may work for high number of small projects. average_monthly_hours plays important role as employee putting too much time for work may have no time to relax and hence will burn out and would like to move on. work accident plays important role as it concern with security of people and that may require people to quit. left is the decision column which will used to predict. promotion_last_5years and salary should have same impact and hencei believe salary should be used and promotion to be dropped. sales is department value and will be interresting to see which department is more impacted. But i do not think it impacts on decision. Salary an important column for employee to quit.

Datasets and Inputs

For this capstone project I have planned to use Human Resource Analytics dataset for anaysis for resources left the company and features that may have played role for moving on. It has below features for evaluation.

- Satisfaction Level
- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Whether they have had a work accident
- Whether they have had a promotion in the last 5 years
- Departments (column sales)
- Salary
- Whether the employee has left

The dataset has total of 14999 rows and 10 columns. I will read the csv file using pandas library. Will anayze the data to check for any null values in any of columns and impact of each column on output. With this model i hope to analyze the employees who are more likely to exit and could do damage control exercise to retain them thus retaining talent and boosting company performance. Target variable (left) is imbalance dataset. It contains 3751 records of employee who have left the company. 11428 records of employee who stayed in the company. Here we are trying to predict employee who can quit. Dataset has been extracted from Kaggle.

Dataset Link: https://www.kaggle.com/ludobenistant/hr-analytics (https://www.kaggle.com/ludobenistant/hr-analytics)

Solution Statement

Company is in a situation where its talented and experience employee are quitting jobs and there might be n number of reasons. HR job is to identify the root cause and take a remedy action. Machine Learning model will help them to identify the people who could quit in future and the reason for quitting the jobs. So that a preventive measure is taken and employee could be retained. I plan to use Classification model of Supervised Learning techniques to learn from history data and when existing employee is pass to the podel it could predict if employee will be moving on and damage control action will be taken. Couple of step wll be taken for preprocessing of data. Will check for null value for feature. If available will fill with mean for numerical column and mode value for catgorical column. Will do one hot encoding for categorical column. There are total of 10 features available. Since all of them will not have equal impact on prediction i will use PCA to dentify the most predictive feature and use those. I will use Ensemble machine learning model(AdaBoost/Gradient Boosting) to understand the employee moving on.Will finalize any one based on performance. I will also use Stochastic Gradient Descent Classifier (SGDC) and SVM to check their f score and if performing better than Ensemble then will use it in fine tune and create model. Since the goal of model is to understand why employees are moving on i will consider more interpretable model than highest prediction accuracy.

Benchmark Model

As a benchmark model i will be using Logistic regression model to predict the Target feature(left). Will use the accuracy and F Beta Score and confusion matrix as benchmark and will strive to get better result in ensemble machine learning model.

Evaluation metrics

Classification model will be evaluated on F Beta score and Accuracy, precision and recall. We can use F-beta score as a metric that considers both precision and recall:

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

In particular, when $\beta = 0.5$, more emphasis is placed on precision. This is called the $F_{0.5}$ score (or F-score for simplicity).

Accuracy measures how often the classifier makes the correct prediction. It is the ratio of the number of correct predictions to the total number of predictions (the number of test data points).

Precision tells us what proportion of messages we classified as left, actually left. It is a ratio of true positives(predicted value for person left and actually person left) to all positives(all the person left company irespective of true or false), in other words it is the ratio of
[True Positives/(True Positives + False Positives)]

Recall(sensitivity) tells us what proportion of employee that left were classified by us as left. It is a ratio of true positives(employee that left, and employee actually left) to all the employee left, in other words it is the ratio of
[True Positives/(True Positives + False Negatives)]

Project Design

- I will import data from csv file using pandas library.
- Once data is available in Dataframe I will use numpy library to calculate the total number of records imported, total number of employee quits the job and total number of employeed chosse to stay. next step will be data preparation for modeling.
- I will convert all the numerical column to one scale and this case it will be log scale, categorical column will be converted to multiple columns with numerical values using one hot encoder technique.
- I will generate graphical representation of different columns with respect to Left column which is decision making column to understand the impact of features on decision making column. Ths will done using matplotlib.pyplot library.
- Once data is prepared i will split dataset into 2 part training and testing dataset with split ratio of 80:20. Will drop the column of testing dataset so that it can be predicted and use to compare actual value with predicted value.
- I will divide the training dataset into 10 different folds to use as Kfold validation helping to minimize risk of overfitting and increasing accuracy of the model. I will first create model using logistice Regression to have a bench mark f score and confusion matrix. Will calculate accuracy as well to check its variation on different mode. F Score will be parameter for judging the performance. I will use Ensemble machine learning(Adaboost/ Gradient Boosting) to train the model.As a backup i will train model based on SVM and Stochastic Gradient Descent Classifier (SGDC) as well to check the performance of other two. I will imrove the model performance using XGBoost to give an extra boost. Here i will use Scikit-learn(sklearn) library to achieve it.
- Further i will apply Grid search technique to fine tune the parameter and increase performance.
- Once model is trained will use test set to predict the value. Accuracy and F Beta score and confusion matrix will be calculated based on predicted value.