Machine Learning Engineer Nanodegree

Capstone Proposal

Ravi Mandava
August 21th, 2018

Proposal

The proposal of the project is to apply machine learning algorithms in ERP (Enterprise Resource Planning) domain for a study on "Employee Attrition Prediction and Analysis". For this requirement, the attempt is to build a model with Supervised Learning algorithms that helps to predict the outcome and analyze the results with the available dataset. This helps human resources managers to give an insight for identifying risks and hence do a better planning.

Domain Background

Enterprise resource planning (ERP) refers to the integrated management of core business processes and automation. The challenge, however, lies in managing this database to get the best use out of the information held within.

Machine learning is all about automating processes that would take more time or difficult to get analytics on dataset with manual methods. The main idea is to organize, transform the available data and provide input to supervised learning algorithms to predict outcome and measure the accuracy. Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset with labeled data) to make predictions. As my current working domain is ERP, it has motivated to apply machine learning techniques in ERP to see the opportunities where business process can be streamlined.

Problem Statement

Employee attrition is a big problem many companies are facing nowadays. On boarding a new employee can become expensive, resources consuming in terms of time, money and sometimes may not even find a perfect replacement. Considering these factors, some of the companies are trying to keep their employees happy and satisfied. One of the ways how to keep employees from leaving is to analyze, why people who left the company decided to do so, predict who could be leaving next and try to take actions for retention. The main objective is to solve the problem by determining the factors that lead to attrition, predict the likeliness of attrition and hence reduce the employees attrition based on results.

Datasets and Inputs

The dataset contains a csv flat file with the HR dataset, containing basic information about the employees (age, daily rate income, years of service, job role, job satisfaction rating, commute distance etc) and whether employee left or stayed with the company. The dataset contains 1470 rows and 35 feature columns with primary feature Attrition (Yes/No). Since this dataset is relatively small, the plan is to make use 80 percent of data for training set and remaining 20 percent data for validation.

The source for dataset is **HR Employee Attrition and Performance**

Identifying the key input features is very important to predict the results with higher accuracy rate. In the dataset chosen, features like employee count, gender may not have much impact on attrition rate. More information can be derived through exploratory data analysis; therefore it is better to avoid these features in the input dataset to reduce model training time.

From the sample dataset, the target Class 'Attrition' is distributed mainly on input features like job role, years of service, working environment, manger relationship, overtime. The data is already preprocessed or captured with rating scales, the percentage of outliners is less and hence don't expect imbalance in classification.

Solution Statement

The most common solutions for labeled datasets can be derived by applying different Supervised Learning algorithms like Random Forest, Decision Tree, and Naives Bayes etc. In order to determine the Employee Attrition rate with the available feature dataset, we implement the above algorithms as solution and predict the employee attrition rate and top factors that influence this.

Benchmark Model

Benchmarking the operational characteristics is as important as evaluating the predictive characteristics of a model. The sample dataset of employee attrition contains both regression features (years of service, age, monthly rate) and classification features (job role, education and department). A simple algorithm ***Decision Tree*** is chosen as benchmark model. The results of this model can be compared to understand how well the new model works.

Evaluation Metrics

The target class 'attrition' should predict correctly with different input features. Accuracy from confusion matrix is an appropriate metric to compare the predicted positive and negative results against true values. The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation:

- TN is the number of correct predictions that an instance is negative,
- FP is the number of incorrect predictions that an instance is positive,
- FN is the number of incorrect of predictions that an instance negative, and
- TP is the number of correct predictions that an instance is positive.

Confusion Matrix	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN
accuracy = (TP +FP)/(TP+FP+FN+TN)		

Project Design

The steps to follow in project design are listed below:

1) ***Feature Extraction, Selection and Transformation*:**
   In this step we extract the important, non-redundant features from raw data and try to get most important features that are contributing to decide the label. The data is transformed, preprocessed or scaled to reduce the outliners.

2) ***Perform Exploratory Data Analysis:***
   In this step the dataset is analyzed to summarize the main characteristics, often with visual methods (Pair Plot, Histograms, scatter plots etc). This is helpful when we have a dataset with many features dataset and an optional step for small datasets.

3) ***Build a predictive model with different techniques:***
   In this step the models like Decision Tree, Random Forest, and Naive Bayes are built by tuning appropriate parameters with common dataset.

4) ***Identity the optimal model through testing and evaluation techniques:***
   In this step the results from each model are passed to the evaluation metrics like accuracy and F1score and identify the optimal model.

5) ***Identify the top features that has more weight to predict the result:***
   In the step based on the EDA (Exploratory Data Analysis) and results from models, factors that have high influence to determine the outputs are rated.

References:

https://towardsdatascience.com/solving-staff-attrition-with-data-3f09af2694cd

https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

https://www.quora.com/How-can-we-use-machine-learning-in-ERP

https://en.wikipedia.org/wiki/Supervised_learning

http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html