PROJECT PROPOSAL

INTRODUCTION TO DATA SCIENCE

Analysis for Employee Retention

Date: 29th November 2017

Submitted to: Prof. Yiheng Liang
Project Team: Bina Maria – Banner ID: 00366107
Saumya Bhatnagar – Banner ID: 00370168
Saurav P. Shrestha – Banner ID: 00369895

Introduction:

Company XYZ has been in industry since a long time. Their business had been increasing quite well over past, however in recent years, there has been a slowdown in terms of growth because their best and most experienced employees leaving prematurely. The VP of the firm is not very happy with the company's best and most experienced employees leaving prematurely wants to find out insights in the company employee data and find out an answer as to know why best and most experienced employees are leaving prematurely.

Objective:

As a first step the VP planned to know the useful insights out of the employee data available. He also wanted his team members to give him a forecast model to predict which employees could be leaving the company, as well as answer to why their best and most experienced employees are leaving prematurely. This will help him plan his next steps to avoid the churn out. We will be designing a script that would contain the following analysis:

• A visualization and distribution (of all the employee relative fields)
• Forecast using different machine learning models
• Comparison among different machine learning models and cross validating through them test and train set
• Find out the factors that most affect the reason why best and most experienced employees are leaving prematurely.
• Give a final prediction model to forecast

Tools required:

RStudio

Actions to be performed:

- Load the dataset in R
    - Use the import functionality or read.csv() command
- Divide data into training set and test set
    - Use the sample() function to split data into training and testing dataset
- Visualize the characteristics of the entire data
    - Use the barplot() function to visualize the characteristics
- Analyze department wise turnout and percentage of employees leaving from each department
    - Use the library plotrix for ploting the data
- Build models using supervised learning algorithms like Decision tree, Random Forest, Naïve Bayes
    - Use packages like:
        - rpart - Recursive Partitioning and Regression Trees. Used for recursive partitioning for classification, regression and survival trees.

- rattle - R Analytic Tool To Learn Easily. It presents statistical and visual summaries of data, transforms data so that it can be readily modelled, builds both unsupervised and supervised machine learning models from the data, presents the performance of models graphically.
- rpart.plot - Used to Plot 'rpart' models.
- RColorBrewer - Provides color schemes for maps (and other graphics)
- caret - classification and regression training. Contains functions to streamline the model training process for complex regression and classification problems.
- e1071 – package used for Naïve Bayes and Support Vector Machine
- randomForest – package for implementing random forest.
- Other packages and functions as and when needed.

DATA TO BE USED:

The dataset being used is Employee_data. It consists of dimensions like:

- Employee 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
- Department
- Salary
- Whether the employee has left