**Module 05 Project**

Employee Churn Classification

Submitted by,

Miss. Shruti Banduji Kamble

College of Prof. Studies, Analytics

**A picture containing indoor

Description automatically generated**

ALY 6020: Predictive Analytics

Prof. Justin Grosz

Assignment Due Date: October 24, 2021

**Introduction:**

This company has gotten a data set from their sister company about elements about people staying or leaving. The dataset has columns like Education, Joining Year, City, Payment Tier, Age, Gender, Ever Benched, Experience in current domain, Leave or not. From these variables we will build models that will help us predicting the employees who will stay or leave the organization.

**Analysis:**

We started analyzing dataset by importing all the required libraries for stats and building the model. Next, we imported the dataset. We have few numerical columns and some categorical columns among the 4653 rows and 9 columns. We ran null value verification test and found zero null values. Hence, we are proceeding for analyzing the dataset. As the business objective is to analyze the employees if leaving or staying, we chose “LeaveOrNot” for visualization and statistics. The histogram shows the data points are not equally distributed over output class. The number (Leave is 1 and not is 0) shows the employees leaving (approx. 1600) the company are less than the employees who are staying (approx. 3100).

Further, we evaluated the relationship of LeaveOrNot with rest of the parameter of dataset. To achieve this, we chose heatmap and the result shows, PaymentTier has the strongest relationship with LeaveOrNot which is logical. ‘ExperienceInCurrentDomain’ and ‘Age’ are the next parameter which has strong relationship with ‘LeaveOrNot’ variable. For preparing data to build the model, we must make sure the data is numerical as numerical data works well with logistic regression and other models. Hence, we have created dummy variables for all the categorical variables. we split the dataset into two sections, one for the dependent variable and the other for the independent variables. After that, we have two datasets, which we must separate into train and test datasets in order to train the model and then test it.

**PART 1:**

The columns that help us determining the employees leave or stay is the dependent variable.

Next, we are separating the independent variables and dependent variable into separate data frames and then split the data into train and test. Now, the data is prepared for building the model.

The first model we are building is Logistic Regression to predict, LeaveOrNot. We need to run the model on this dataset to see which factors influencing the ‘LeaveOrNot’ of the company. We're ready to run the model now that we have the train and test data. The logistic regression shows 73.22% accuracy and precision rate is 69.11%. The variable ‘JoiningYear’, ‘PaymentTier’, ‘Age’, ‘ExperienceInCurrentDomain’ are the most significant variable which are influencing employees to leave the company or not. PaymentTier with p-value of 3.290929e-08 and coefficient value -0.337553 and ExperienceInCurrentDomain with p-value of 3.696253e-02 and co-efficient value of -0.045623. This concludes that the PaymentTier and ExperienceInCurrentDomain increases the probability of an employee leaving the company decreases.

**PART 2:**

The next model we build is Random Forest Model. To produce a more exact and dependable conjecture, Random Forest makes numerous decision trees and mixes them together. Arbitrary timberland has the advantage of having the option to tackle both characterization and relapse issues. At the point when we have an enormous dataset and interpretability is certifiably not a key concern, Random Forest is a decent decision.

The accuracy of the Random Forest model is 82.45% with 200 estimators. The precision rates for the employees leaving the company is 65% and for the employees staying with the company is 74%. The result took 7.081031799316406e-05 seconds.

**PART 3:**

The next model is Neural Network. Neural Network model help us recognizing patterns and work by and work by grouping or classifying raw data. We ran the model and from the result, we have got 81.66% accuracy at 30000 iterations and at 20000 iterations we have received 81.73% accuracy. We can see that the model with different iterations is the accuracy of the model is remains the same meaning the model of the neural network is good.

**Conclusion:**

By working on analysis, we can conclude that PaymentTier and ExperienceInCurrentDomain which are inversely influencing the employees. The accuracy score for the logistic regression is 73.22% and precision rate is 69.11%. The precision of the model is also good for both values. Next model we ran is Random Forest where the accuracy of the model is 82.45% and the precision rate is 65%. The precision of the random forest model is not performing well even though the accuracy of the model is good we can conclude that we don’t have to select random forest for this problem. Next the neural network with multiple iterations the accuracy of the model remained almost same at 81.66%, 81.73% and 82.38% for 10000, 20000 and 1000 iterations respectively. The logistic regression model’s accuracy is 73.22% and neural networks accuracy 81.66 % even though the difference between the models is not too much but the sophistication of neural networks model than logistic regression makes it easier for us to recommend neural network to the school to understand the employee behavior.

**Reference:**

Chhaya Bakshi (June 8, 2020): Random Forest Regression retrieved from <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>

Chris Nicholson: A Beginner's Guide to Neural Networks and Deep Learning retrieved from <https://wiki.pathmind.com/neural-network>