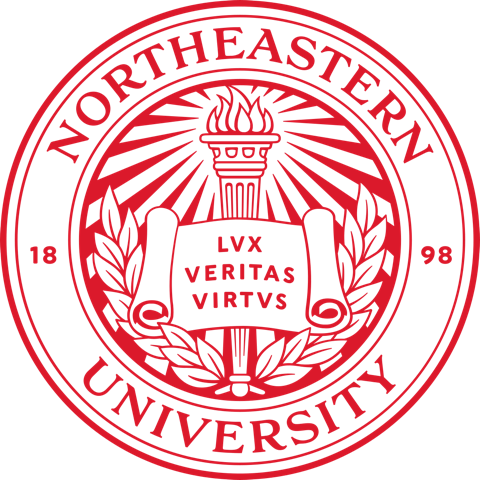
**Module 5 Project**

**Employee Churn Classification**

Sunil Raj Thota



College of Professional Studies, Northeastern University ALY 6020 – Predictive Analytics

Prof. Justin Grosz

December 07, 2021

**Introduction:**

A business wants to learn more about its employees. This company received a data collection regarding persons staying and leaving from its sister company. They want to see if we can construct a strategy that can reliably forecast who will stay or leave now that they've recruited a data analyst. Understanding why and when employees are most likely to quit might help you take steps to increase employee retention and possibly plan for new hires. Employee turnover is a costly issue for businesses. The true cost of replacing an employee might be fairly high in many cases.

**Data Quality, Cleansing, Preprocessing, and Exploratory Data Analysis:**

To begin, we must cleanse the data and identify any outliers (if any) to ensure that we have high-quality data for the model. We will develop a logistic regression, random forest, and neural network models to estimate whether the employees are leaving or not based on the numerous features in the dataset using optimization approaches to see if we can improve the model's accuracy and compare the models, as stated in the question. After evaluating the findings, we will advise the company on which factors the employees will stay or go. We've imported all of the packages and libraries we'll need for our initial data exploration. Using the scikit-learn to package, you can create, evaluate, and tune various classification models.

There are 4653 records and 9 columns in the dataset. To get started, we've imported all of the packages that are required for undertaking model analysis. Pandas, NumPy, matplotlib, seaborn, train test split, Logistic Regression, Random Forest Classifier, MLP Classifier, and preprocessing are all examples of Python libraries. We tested for missing values in the dataset after putting it into a Python environment for further analysis. Missing values must be taken into account because they may have an impact on our analysis and AI models.

There are no null values detected in the dataset. There are 4 categorical variables namely Education, City, Gender, and EverBenched. Each of these attributes has been classified into 2 or more categories. We investigate the distribution of Education, Age, Gender, City, PaymentTier, EverBenched, ExperienceInCurrentDomain to the LeaveOrNot more fully because several classification algorithms rely on a logit relationship between features and target. To find patterns, we'll look for feature correlations first. I have label encoded the categorical variables and transformed them using the fit\_transform() method.

**Part 1:**

**Logistic Regression Model:**

We will be using the Logit function to fit the model with the necessary variables to see the p values and other statistics of each column. I also went ahead and visualized the correlation between the variables using a corr heatmap. Used liblinear as the solver in the Logistic Regression method observed the model fit. And then I have calculated the Accuracy, Precision, Recall, and F1-Scores. The ratio of correctly classified subjects to the total number of subjects is known as accuracy. The ratio of accurately +ve labeled to all +ve labeled is known as precision. A recall is the proportion of those who are in reality to those who are appropriately +ve classified. Precision and recall are both taken into account while calculating the F1 Score. It's the precision and recall's average.

The top 3 significant variables are EverBenched, City and Education are mainly driving the employees to leave. From this, we can understand that there is more impact by the EverBenched because an employee is made to sit idle without any work/ project/ task. As there will be more demand in the cosmopolitan cities, employees tend to move and look for newer and wider opportunities. It also mostly depends on whether the education of an employee is highly skilled or not. We can more focus on these 3 targets and find and strategize accordingly and make employees satisfied to retain and there is a lot of room for growth. Provide personalized data-driven metrics, performance reports, and retention plans.

**Part 2:**

**Random Forests Model:**

Random Forest is an ensemble learning-based supervised machine learning technique. I built two Random Forest Classifier models in this project to predict whether the employees are leaving or not. With the number of decision trees in the model, the expected accuracy rises. I presented the feature selection process by utilizing the Random Forest model to locate only the most significant features, then rebuilding the model with these features to evaluate how accurate it is. max depth: The total number of splits for all of the forest's trees. bootstrap: Indicates whether or not bootstrap samples should be used while creating trees. max features: The maximum number of features that will be used in node splitting – this is the key difference between bagging trees and random forest as I previously discussed. In most cases, you want a value less than p, where p is the total number of features in your data set. criterion: This is the metric used to evaluate the decision trees' stopping criteria.

Joining Year, Age, and City are the most significant variables in this model. From this, we can understand that there is more impact by these respective attributes where employees are more inclined to leave the company. It depends on their joining year in the company and the years that they had worked. And it also depends on the Age of a particular employee (25 - 45). We need to know more details on the city because Employees tend to move and hunt for greater and bigger prospects in developed areas because there will be more in demand and make a decision accordingly.

**Part 3:**

**Neural Networks Model:**

The perceptron is the first step in building a neural network. In simple words, the perceptron takes inputs, multiplies them by some weights, and then sends them through an activation function (such as logistic, ReLU, tanh, or identity) to get an output. MLP Classifier stands for Multi-layer Perceptron Classifier, which is linked to a Neural Network by its name. Unlike other classification methods such as Support Vectors or Naive Bayes Classifier, MLP Classifier does classification using an underlying Neural Network. However, MLP Classifier is identical to Scikit-learn's classification algorithms in that it requires no more effort to implement than Support Vectors, Naive Bayes, or any other Scikit-Learn classifier. hidden layer sizes: The hidden layer sizes option allows us to specify the number of layers and nodes in the Neural Network Classifier. Each tuple member reflects the number of nodes at the ith position, where i is the tuple's index.

The total number of hidden layers in the network is thus represented by the length of the tuple. max iter is a variable that represents the number of epochs. activation: The hidden layers' activation function. solver: The algorithm for weight optimization across nodes is specified by this parameter. I have used the ‘lbfgs’ solver. In terms of both training time and validation score, the default solver 'adam' performs admirably on relatively big datasets (with thousands of training samples or more). 'lbfgs', on the other hand, can converge faster and perform better for small datasets. random state: This parameter allows you to specify a seed that will allow you to reproduce the same results.

After running the model, I have achieved an accuracy of 0.66 with a ROC AUC Score of 0.5. Because it is difficult to regulate the training of a multi-layer perceptron (MLP) classifier, its performance on unseen patterns is unpredictable. One of the major issues with training an MLP classifier is overtraining. The partial derivatives of the loss function to the model parameters are computed at each time step to update the parameters, therefore MLP Classifier trains iteratively. It can also have a regularization term added to the loss function to prevent overfitting by shrinking model parameters. This method works with floating-point data encoded as dense NumPy arrays or sparse SciPy arrays. MLP requires tuning a number of hyperparameters such as the number of hidden neurons, layers, and iterations. MLP is sensitive to feature scaling.

**Part 4:**

**Comparisons, Findings, and Recommendations:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Speed** | **Accuracy** | **Precision** | **Recall** | **MSE** | **ROC AUC Score** |
| Logistic Regression | 0:00:00.024 | 0.71 | 0.64 | 0.34 | 0.29 | 0.67 |
| Random Forest Model | 0:00:00.099 | 0.84 | 0.81 | 0.69 | 0.16 | 0.98 |
| Neural Networks Model | 0:00:00.111 | 0.66 | 0.0 | 0.0 | 0.34 | 0.50 |

From the above models, Joining Year and City are the 2 majorly contributing parameters to the prediction of whether the employees are leaving or not. So, I would recommend this company to focus more on the data whether the employees are satisfied or not, their personalized performance, quality of projects, and cities they are living and working. From the above results, I would recommend Random Forest Classifier Model as the best model to go with because of its optimal Speed, Accuracy, Recall, Precision, and MSE Scores.

When coming to City, employees who reside further away from the office are more likely to leave. As a result, efforts should be made to assist in the form of business transportation or Transportation Allowance for groups of employees departing the same location. Employees should not be initially screened based on their home location because this would be considered discriminatory as long as they arrive at work on time every day. Employees with more experience are less likely to leave. Employees with more than 4-5 years of experience should be flagged as potentially leaving at a higher rate.

**Conclusion:**

The company has to find and strategize the people management with proper data insights where they have a lot of room for growth. As obtaining the results, the company has to change its management strategy and implement a detailed data-driven decision-making analysis and recommendations to the employees. For each Risk Category group, a strategic "Retention Plan" should be written out. Face-to-face meetings between an HR representative and employees can be begun for medium- and high-risk employees to discuss work circumstances, in addition to the prescribed measures for each feature described above. A meeting with those employees' managers would also allow them to discuss the team's working environment and whether any efforts can be done to enhance it.

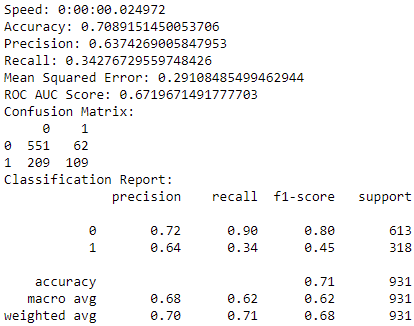
**References:**

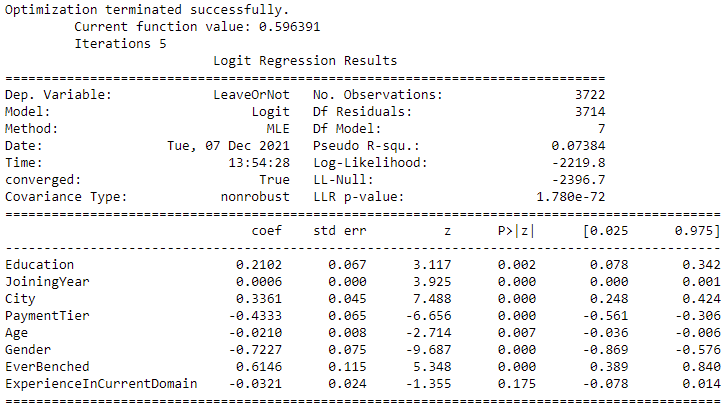
Amal Nair. (June 20, 20)19. A Beginner’s Guide To Scikit-Learn’s MLPClassifier. *Analytics India Mag*.

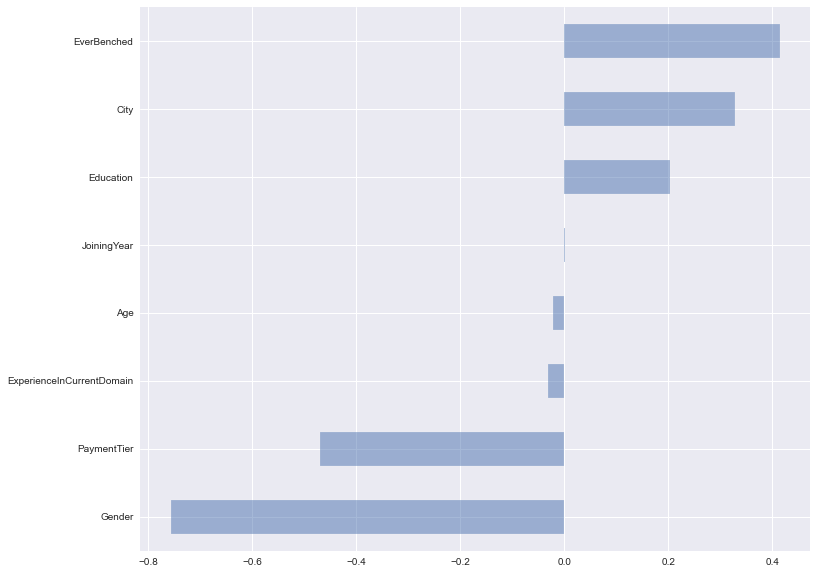
Retrieved from https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/

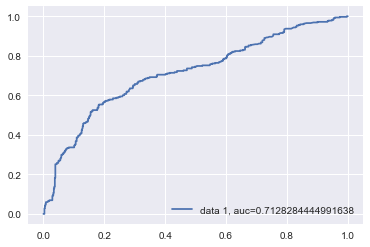
**Appendix:**

**Figure 1: Logistic Regression Results**

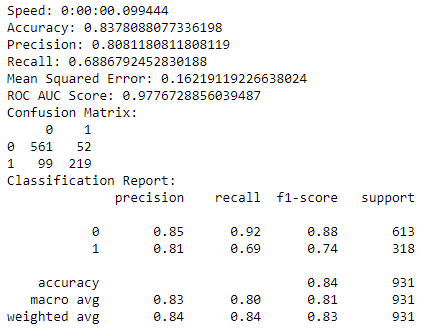
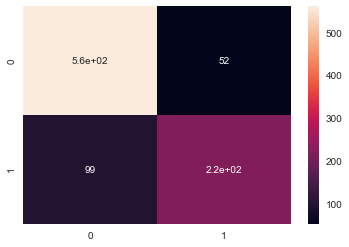
 

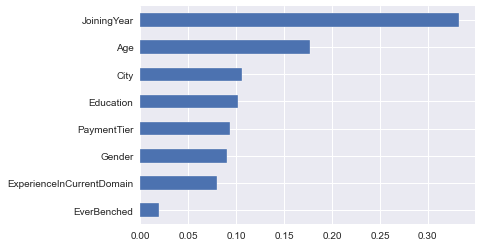




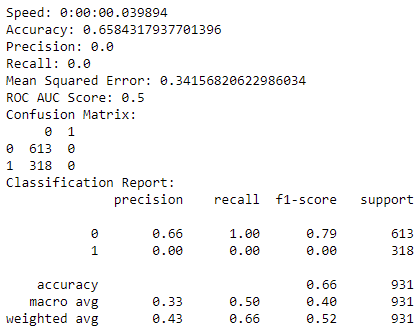
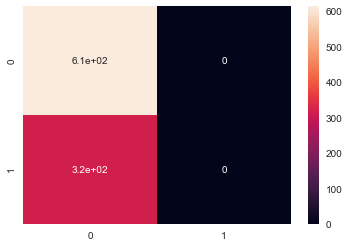


**Figure 2: Random Forest Results**

** **

****

**Figure 3: Neural Networks Results**

** **