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.

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.

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.

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. 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 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.

Part 2:

Random Forests Model:

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. In most cases, you want a value less than p, where p is the total number of features in your data set.

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:

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.

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).

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. 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
Logistic Regression	0:00:00.024	0.71	0.64	0.34
Random Forest Model	0:00:00.099	0.84	0.81	0.69
Neural Networks Model	0:00:00.111	0.66	0.0	0.0

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.

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.

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

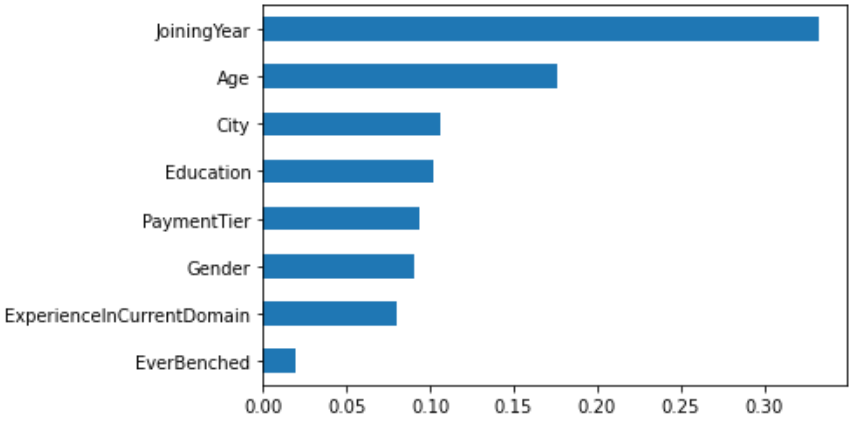
Optimization terminated successfully.
Current function value: 0.596391
Iterations 5

Logit Regression Results

Dep. Variable:	LeaveOrNot	No. Observations:	3722
Model:	Logit	Df Residuals:	3714
Method:	MLE	Df Model:	7
Date:	Sat, 11 Dec 2021	Pseudo R-squ.:	0.07384
Time:	18:05:23	Log-Likelihood:	-2219.8
converged:	True	LL-Null:	-2396.7
Covariance Type:	nonrobust	LLR p-value:	1.780e-72

	coef	std err	z	P>|z|	[0.025	0.975]
Education	0.2102	0.067	3.117	0.002	0.078	0.342
JoiningYear	0.0006	0.000	3.925	0.000	0.000	0.001
City	0.3361	0.045	7.488	0.000	0.248	0.424
PaymentTier	-0.4333	0.065	-6.656	0.000	-0.561	-0.306
Age	-0.0210	0.008	-2.714	0.007	-0.036	-0.006
Gender	-0.7227	0.075	-9.687	0.000	-0.869	-0.576
EverBenched	0.6146	0.115	5.348	0.000	0.389	0.840
ExperienceInCurrentDomain	-0.0321	0.024	-1.355	0.175	-0.078	0.014

Figure 2: Random Forest Results


JoiningYear
Age
City
Education
PaymentTier
Gender
ExperienceInCurrentDomain
EverBenched
0.00
0.05
0.10
0.15
0.20
0.25
0.30