Non-Linear Model as a tool for Human Resources (HR) Predictive Analytics

SPE ML Bootcamp Team 3

Erica Esatyana, Christian Woods, Mark Yang, and Sabayasachi Prakash

1. Problem Introduction

HR predictive analytics offers an effective and efficient way to identify the employees that have a huge chance to get promoted based on several factors. It is indispensable to constrain HR while selecting a candidate objectively. A large company or organization that has hundreds to thousands of employees will be beneficial using this analytical approach specifically in a short amount of time to avoid any delay in transition that can affect the promotion cycle. Our company will justify the use of SVM or neural networks in predicting whether the employees get promoted or not.

2. Dataset

Our dataset was attained from the open-source platform Kaggle. It consists of 13 input features and 1 output feature. The total dataset is 54,800 2409 4124.

3. Features and Processing

We first explored the dataset by searching for missing values. “Education” and “previous year rating” were found to have 2,409 and 4,124 missing values, respectively. These missing values were handled (NaN) by dropping them. We hypothesized that dripping these values should not affect our results as the dataset is quite large. Some features such as employee_id, department, region, and recruitment_channel were dropped because we observed those might not contribute to the decision whether employees get promoted or not. The input features like gender and education were mapped into categorical and ordinal variables, respectively. Data preprocessing was done using sklearn minmaxscaler(). The data set was found to be heavily imbalanced with 8% of promotions.

4. Models and Techniques

Multiple classification algorithms were implemented for this project to compare their capability of handling the imbalanced dataset.

- Support Vector Machines (SVM) are a class of machine learning methods that can be used for classification as well as regression problems. They are based on the idea of determining a hyperplane which most appropriately separates the data into its class.
- Neural network Multi-layer Perceptron (MLP) is a supervised learning algorithm in which the algorithm can learn a function from the dataset. It is capable of learning a non-linear function approximator for classification when given a set of features and a target.
- Logistic Regression is a classification algorithm used to designate observations to discrete classes based on probability. It uses the sigmoid cost function to map predicted values to probabilities.

- KMeans is an unsupervised machine learning algorithm that uses clustering to divide observations into subgroups based on their similarity.
- K nearest neighbors (KNN) classification is a supervised machine learning algorithm that assigns observations to groups based on similarity measures by a distance function.
- Naive Bayes Classifier is a machine learning model used for classification based on Bayes theorem.
- Random forest classification is an ensemble technique that constructs decision trees and aggregates the votes from different decision trees to determine the output classification.

Model	kernel	Accuracy	Precision	Recall	F1 score
SVC	Poly, d2	0.91	0.84	0.08	0.15
Neural Network	MLP	0.84	0.06	0.05	0.06
Logistic regression	----	0.91	0.67	0.06	0.12
Kmeans	----	0.67	0.17	0.70	0.27
KNN	----	0.91	0.53	0.20	0.29
Naive Bayes	----	0.91	0.48	0.13	0.21
Random Forest	----	0.91	0.52	0.23	0.32

5. Results and Discussion

. All of the models, excluding the K-means model, had decent accuracy statistics. Unfortunately, accuracy can be a very misleading statistic for imbalanced datasets. Approximately 92% of people in our dataset did not receive a promotion. This means that a model could have 92% accuracy if the model were to only predict that every individual did not receive a promotion. A better metric for assessing the quality of the model predictions is the F1-score. It is calculated using precision (positive predictive value) and recall (true positive rate), which makes it a much more robust statistical measure. The imbalanced dataset resulted in many false negative predictions leading to low recall values for all models except for Kmeans. These low recall values greatly diminish the F1-scores as well. The best model based on the criteria of the highest F1-score is Random Forest. Despite not having the highest precision between all models, the improved recall value resulted in a higher F1-score. Nevertheless, none of these classifiers have done a great job of classifying the imbalanced class.

6. Conclusion

The imbalance classification dataset is one of the most challenging datasets to work with. Many standard classification algorithms have a bias toward the majority class. The features of the minority class are treated as noise in the dataset leading to their dismissal. Some algorithms do worse than just having an algorithm output of one class (‘0’). Due to the time constraints of the project, we were unable to implement some of the common techniques for dealing with imbalanced

data such as resampling and ensemble techniques. From the models we implemented, Random Forest produced the best F1-score by improving recall with a reasonable amount of Precision.