SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Classification Project

Machine Learning 2

Assignment 1

Mrad Azoury

The HR analytics dataset comprises of explanatory variables of around 15k employees of a large company. The main goal of this classification problem is to model the probability of employee attrition (both the ones who left on their own or the ones who got fired). Furthermore, the model will allow us to understand which are the most important variables and the ones that need to be addressed right away.

1- Data Loading and Data preparation: (Exploratory Data Analysis)

After loading the data, an exploratory data analysis was done to provide insights on predictors. In addition, this analysis provides data structure as well and both, in accordance, will guide the feature engineering work. First of all, all the libraries used were imported and the function to read the prepared dataset and the raw dataset was defined. Another function was defined to fix the types of the three following variables: Work_accident, promotion_last_5years and left, converting them to categorical variables. Both variables; salary and sales were dummified, sales were divided based on the different departments that we have resulting in 8 new columns (with the different departments) and salary was divided into three levels low, medium and high resulting in three new columns.

The function feature_skewness was defined which gets the features that need to be fixed and based on that these features or variables will be fixed by running the second function that was defined; fix_skewness. Same for the scaling (feature_scaling) in order to rescale the numerical variables. No outliers were detected while preparing the dataset. In addition, the correlation was checked using spearman’s function and no correlation was detected.

2- Baseline model:

A basic model was run over the raw dataset to check what is the model performance on the classification task that we want to achieve using logistic regression. As a test, the dataset was split into 80% (training set) and 20% (test set). The accuracy score for the baseline model is 0.795.

3- Feature engineering:

In order to understand what can be done to increase the accuracy of the model, some variables were plotted against the employee attrition to check for their effect on the churn rate.

Feature creation:

1- Total hours; creates a new column of the time spent in the company (in years) multiplied by 12 which is multiplied by average monthly hours and then plotting it in relation to employee attrition. It might be relevant to see if the churn rate is affected by the total hours spent by an employee within the company. This feature was then binned according to the graph distribution (function: bin_total_hours).

2- To make sense out of the data that we have, Relative_working_hours_dept feature was created which is calculated by:
(average monthly hours – average of the average monthly hours for the employees within the same department) / standard deviation of the average monthly hours for the employees working within the same department.
Generally, employees compare their working hours to the employees that work within the same department, hence this feature creation allows us to get more insights about the employee attrition within the different departments.

3- Similarly, Relative_last_evaluation_dept feature was created which was calculated by the following formula:
(Last_evaluation – average of the last evaluation for the employees within the same department) / standard deviation of the last evaluation for the employees working within the same department.
This feature gives us insight on the employee performance within every single department and the effect it has on the churn rate.

4- Moreover, relative_satisfaction_dept feature was added following the same logic above and it was calculated accordingly:
(Satisfaction_level – average of the satisfaction level for the employees within the same department) / standard deviation of the satisfaction level for the employees within the same department.
Normally, within the same department, people compare themselves to their fellow colleagues especially when it comes to their satisfaction level, for example if an employee is unhappy in the HR department (which has relatively a high average satisfaction level) this employee might tend to leave the company more than an employee that has the same satisfaction level in another department which has a lower average satisfaction level.

These features were then plotted separately in relation to the employee attrition. And then, these features were binned based on the similar behavior seen in the plots (using the functions binRelativeAvgHrs, binRelativeSatLevel and binRelativeLastEval).

To further improve the accuracy of the model, the following features were grouped and were defined by these functions: group_time, group_project and group_depts. According to the plots of these features, each one was grouped subsequently:

- The 2, 7, 8 and 10 years of time spent within the company were merged in the same group because they tend to have a very negligible level churn which were then dummified.

- The projects 2, 6 and 7 were grouped into one category and then dummified.
- The departments (product management, R&D, marketing, accounting, management and HR) were merged into one category because they are small departments and have the same behavior or frequency of churn. And then they were dummified.

Since it seemed that the satisfaction level is one of the most important features to determine the churn rate, it was binned after analyzing its graph using the function binSatisfactionLevel.

Feature Learning:

- Genetic Programming (Function: Genetic_P) was used to create new features. It takes into consideration the most important features and then tries to find hidden relationships between them using basic mathematical functions. After applying this, 15 new columns were generated.

Before moving to feature selection, in order to test for the features, a cross validation method was used in a pipeline function to make sure that only the features that do not decrease the accuracy are applied on the dataset depending on the average score of the number of splits taken. The whole dataset was used here just to make sure that the features are valuable, **however the model trained wasn't implemented here yet.**

```
- New mean Score (relative_satisfaction_dept): 0.7901 [diff: -0.0014] [Accepted]
- New mean Score (relative_working_hours_dept): 0.7919 [diff: 0.0018] [Accepted]
- New mean Score (relative_last_evaluation_dept): 0.7917 [diff: -0.0002] [Accepted]
- New mean Score (total_hours): 0.7912 [diff: -0.0005] [Accepted]
- New mean Score (bin_total_hours): 0.8787 [diff: 0.0875] [Accepted]
- New mean Score (binRelativeAvgHrs): 0.8961 [diff: 0.0174] [Accepted]
- New mean Score (binRelativeSatLevel): 0.9301 [diff: 0.0340] [Accepted]
- New mean Score (binRelativeLastEval): 0.9386 [diff: 0.0085] [Accepted]
- New mean Score (group_time): 0.9473 [diff: 0.0087] [Accepted]
- New mean Score (binSatisfactionLevel): 0.9610 [diff: 0.0137] [Accepted]
- New mean Score (group_depts): 0.9607 [diff: -0.0003] [Accepted]
- New mean Score (group_project): 0.9623 [diff: 0.0016] [Accepted]
- New mean Score (feature_normalizing): 0.9622 [diff: -0.0001] [Accepted]
-Number of features created out of genetic programing: (14999, 12)
- New mean Score (Genetic_P): 0.9713 [diff: 0.0091] [Accepted]
- New mean Score (PCA_F): 0.9713 [diff: 0.0000] [Accepted]
Dataset shape AFTER feature engineering: (14999, 63)
```

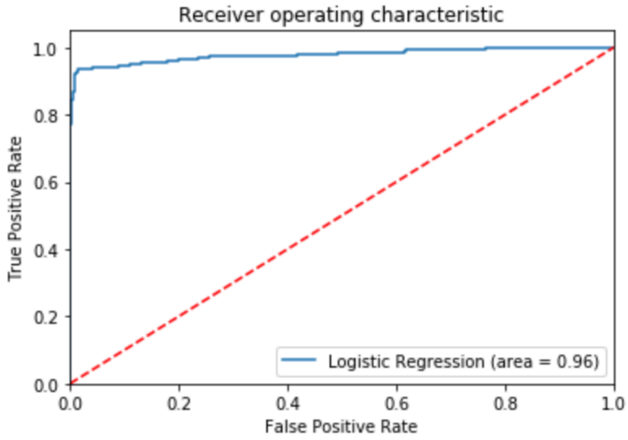
Feature Selection:

- PCA (Principal Component Analysis) was used to convert the set of observations from the numerical variables and remove any possible correlation that may have been found or created from the above features selected.
- RFE (Recursive Feature Elimination): after both feature creation and feature learning, the number of features reached 66. There is no doubt that not all of them have the same importance. So, in order to get the optimal number of features needed and the most relevant ones to get the highest accuracy score possible, RFE was implemented while cross validating using k-fold of 10. It turned out that an optimal number of 30 features was selected.

4- Final metric:

Using the selected features, a holdout test set of 20% was taken aside. The other 80% was used to train multiple models using k-fold cross validation in order to minimize overfitting and to make sure that the model is not biased. The final model chosen was the one that obtained the highest accuracy score across the different k splits.

This model is finally used to predict the target variable of the holdout test set which lead to an accuracy score of 0.9723.



```
[[2246   29]
 [  54  671]]
             precision    recall  f1-score   support

      False       0.98      0.99      0.98      2275
       True       0.96      0.93      0.94       725

avg / total       0.97      0.97      0.97      3000

The final accuracy score is : 0.9723333333333334
```