Predicting Healthcare Employee Attrition

Brief by Jingcheng Li (2034306)

Introduction - The healthcare industry is one of the many industries that are experiencing staff turnover today due to threats such as the COVID-19. In order to maintain the stability of the staff while effectively preventing the gap caused by the loss of staff, it is necessary to collect the information of the practitioners through big data technology and discover the potential pitfalls in advance through data analysis. In addition, such method can help technicians understand the root causes of turnover, so that they can take steps to reverse the trend.

Methodology – The specific steps of data preprocessing carried out in this project are as follows: (1) **Delete null values.** However, with the 'isnull' function I did not find any null values in the dataset. (2) **Delete unrelevant columns.** Some columns in the dataset are the same for all samples and do not affect the model's judgment of 'Attrition'. Therefore, these columns were deleted. (3) **Replace unnormal name.** After observing the dataset, I found that there are two conflicting data in the 'JobRole' column: 'Admin' and 'Administrative'. As a result, I replaced all 'Admin' with the former in order to avoid unnecessary classification errors. (4) **Delete Outliers.** To find the columns with possible outliers, I first calculated the extreme values and average values of all the numerical columns. According to the output, I used the Z-score value to identify samples that exceed the threshold and removed them. (5) **Calculate average satisfaction.** Considering that there are three attributes that are all related to satisfaction, I reduced the total number of attributes by averaging them into a new feature 'AvgSatisfaction' and deleting the original three. (6) **Normalization.** In order to prevent the data from varying too much due to different units, I normalized eleven important attributes with 'MinMaxScaler'. (7) **Data type conversion.** Since columns 'Attrition', 'Gender' and 'OverTime' in the table are formatted as strings, I converted them to floating-point or binary data for facilitating the computation of the correlation matrix and the training of the model. (8) **Data visualization and features selection.** Temporarily removing some of the nominal attributes, I used the correlation matrix for all attributes of numeric type. The attributes that are positively and negatively correlated with the Attrition attribute can be clearly seen in the figure [Fig 1]. For those nominal attributes, I used the 'catplot' function to generate histograms to visualize their relationship with 'Attrition'. Ultimately, I chose 15 attributes with the highest degree of association with the target to be extracted as features.

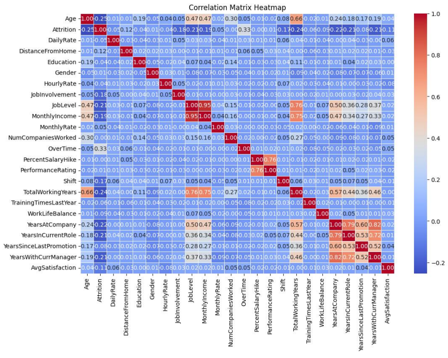
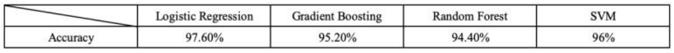


Fig 1. Correlation Matrix for different attributes

During the training and testing phase, I used a logistic regression model. This model finds a value between 0 and 1 for the input by using a Sigmoid function and classifies the input by comparing it with a threshold value. During training, the model utilizes a cross-entropy loss function to adjust the magnitude of the parameters, which ultimately leads to the solution of the binary classification problem.

Results - In addition to using Logistic Regression, I also used other classifiers such as Gradient Boosting, Random Forest and SVM for a side-by-side comparison. However, none of them ended up testing as well as logistic regression. The specific test results are as follows:



	Logistic Regression	Gradient Boosting	Random Forest	SVM
Accuracy	97.60%	95.20%	94.40%	96%

Fig 2 . Accuracy of different classifiers

Discussion

- Pro: This method effectively predicts potential turnover and reduces time and resource consumption in resolving such issues.
- Con: In practice it is difficult to find associations between sample attributes, which makes feature extraction difficult.
- As a contributor, I think this technology should be focused on improving its robustness to rapidly changing data. With the introduction of various efficient data processing methods, I believe that the application prospects of big data technology in fields such as sales and manufacturing are immeasurable.

Conclusion - This project uses a large number of data preprocessing methods. Compared to the original data, the processed attributes show a high degree of correlation with the target. In addition, in the comparison of models, I found that logistic regression model has advantages in the classification of binary classification data sets.