It is group work by Xiaojia He, Qi Sun, Manling Yang.

1. Introduction

For the final project, we'll do research on employee attrition. The data used for this project was collected from the personnel records of employees in IBM. We downloaded this dataset from Kaggle. Although there are many works done by using this dataset, we believe if we use different sets of features, use different methods to handle and transform the data, and use different machine learning algorithms and evaluation methods, we'll get different results.

Attrition is one of the important issues in an organization. Stovel, M. and Bontis, N. (2002) draw attention to controlling attrition, they state that the value of employees to an organization is a very crucial element in the success of the organization. When employees leave an organization, they carry with them invaluable knowledge which is often the source of competitive advantage for the business. Attrition is defined as the normal and uncontrollable reduction of a workforce because of retirement, death, sickness, and relocation. The employee attrition is often unpredictable and can leave gaps in an organization. If organizations know why their employees leave, they can develop effective strategies for employee retention. The main purposes of this study are to find out why employees left the organization and to develop a model that can predict how likely the employee quit the job.

There are many reasons why employees leave the organization.

Based on the research of Firth et al (2007), intention to quit is largely influenced by job dissatisfaction, lack of commitment to the organization and feelings of stress. The work overload and job ambiguity which are the factors that trigger the chain of psychological states that lead to intention to quit. Reducing attrition can save organizations the considerable financial cost and
effort involved in the recruitment, induction and training of replacement staff.

Griffeth et al. (2000) concludes that the predictors of employee turnover include job satisfaction, organizational commitment, job search, comparison of alternatives, withdrawal cognitions, and quit intentions. Also, job performance can foreshadow turnover. The pay and pay-related variables have a significant effect on employee turnover. The results of the gender-turnover correlation indicates that women turnover rate is similar to that of men. The gender age-turnover relationship demonstrates that women are more likely to remain as they age than are men.

Louis (1980) states that attrition takes place because new employees compare their actual experience with their past work experiences. Past work experience plays a significant role in taking decisions to quit in case the new worker’s expectations are not met.

All the researches mentioned above have been completed over ten years ago. In this study, we’ll examine if the factors mentioned in above literature can also result in employee attrition in IBM. We’ll also develop a model to predict the employee attrition. The outcome of this research can be utilized for redesigning the HR policies and practices and take corrective actions to reduce the attrition rate.

References:

Stovel, M. and Bontis, N. (2002), "Voluntary turnover: knowledge management – friend or foe?", *Journal of Intellectual Capital*, Vol. 3 No. 3, pp. 303-322.
https://pdfs.semanticscholar.org/1856/1de816a436871587869169369dfe086ecb6c.pdf

Firth, L., Mellor, D.J., Moore, K.A. and Loquet, C. (2004), "How can managers reduce employee intention to quit?", *Journal of Managerial Psychology*, Vol. 19 No. 2, pp. 170-187.
https://pdfs.semanticscholar.org/4ffe/89346438e0d97e4241ca76779e153247d91a.pdf?_ga=2.41973861.1951415783.1594600481-158954980.1590280415

Griffeth RW, Hom PW, Gaertner S (2000). "A meta-analysis of antecedents and correlates of employee turnover: update, moderator tests, and research implications for the next millennium", J. Manage. 26 (3): 463-88.
https://d1wqtxts1xzle7.cloudfront.net/42653895/A_Meta-Analysis_of_Antecedents_and_Corre20160213-22360-fjbwz8.pdf?1455393954=&response-content-disposition=inline%3B+filename%3DA_meta-analysis_of_antecedents_and_corre.pdf&Expires=1594604746&Signature=F6EiCVoswSyNtj9F0jOg-SYIIiP5ZdRS8zc16PjlQyPnJ4qpB07~aLkaINRqyFkmqi13NWogN0gaWnyawe6zuXuelwmMHQICvdZxsaiqO2AjkpDA1vZ8xkQKlEhwdtEvzCO-9NWLXcBDwHYAcHm1R1IpDHaMQ1YzmoV1FKN2qVd8ip0Wa4UGIfIb1VN5lNnV9Q-z0Nx-FlpoYks8BEnO-GvTbfZENTAMvzAVNaD59FOttbx1NH6AXUGfZCdYCKHYQolMzluDEybDEH9jbLWjSg2CHFK9lhKNvSu~Zz1XDFR-2Tnm0~stCwtX6KNX66KNWOig30lasiq9ynho-Vqnvw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Louis, M.R. (1980), "Surprise and sense-making: what newcomers experience in entering unfamiliar organizational settings", Administrative Science Quarterly, Vol. 25 No.2, pp.226-51.
http://pdfs.semanticscholar.org/80f2/e6d17eecc0f05dffed9b8c4d740d77373f25.pdf

2. **Research Questions**
   Provide a single succinct sentence describing each of your research questions. Then provide a paragraph or two explainings how the results of your research might be used/implemented in the "real world"

The following research questions will guide the study.

1. To what extent do independent variables predict the employee attrition?

   The results of this can help people know how well the model we build. In the project, we will create several models to find the best one by the confusion metrics report when the response variable is categorical data. In the real world, this result tells us if the models are meaningful. For example, if the accuracy of the model is less than the null error rate, this model is not meaningful. We should create the new model or select the other features.

2. What are the relationships between attrition and other variables?

   According to the correlations between one and other variables, we can directly find out if there is any relationship between them. If yes, it is positive or negative. Moreover, if there are high correlations, they may have multicollinearity which must be judged if we need to delete it.

3. Which features are most contributed to the response variable?

   The results are not only beneficial for us to reduce the model overfitting problems but also help us discover better features to predict the employees’ attrition. In the real world, the HR department can increase the performance of the most contributed features to decrease the employees’ leaving.

4. What is the best model to predict employees' attrition?

   The best model can generate the best predictions of employees' attrition based on different metrics. It is a useful way for us to discover what features are the main influence on employees' attrition. In the real world, HR departments can figure out which employees want to quit the job. Other than that, HR may improve the main features which belong to the best model for employees to keep the highly qualified staff.

5. What is the percentage of attrition in each independent variable?

   We can directly see the employees' attrition differences and comparisons in each categorical independent variable to understand what has higher or lower numbers of employees' attrition. We will know the basic employees' attrition situations in different fields. According to the result, we can discover if the feature has a significant impact on attrition. If there is a significant effect, we should consider it to be a part of the models when we do feature selection.

3. Data to be used

Clearly identify the sources of your data and explain the methods you will use to collect the data from those sources, e.g., "Data will need to be collected from this source via scraping of a web page..", etc.

We download the dataset about IBM HR analytics employees' attrition and performance from the Kaggle website https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv . After we download the data, we will upload it to Github. We will read the data from Github when we begin to work on the project in code.

4. Approach

The main process of the project that we plan

Step 1: Load data into the data frame from GitHub

Step 2: Exploratory Data Analysis

Step 3: Data Cleaning and preparation

Step 4: Feature Engineering and Feature Selection

Step 5: Logistic Regression Creation

Step 6: Model Evaluation

Step 7: Conclusion

>For **Step 2**, the main components of EDA include:
>1. Data exploration
2. Check numerical and categorical columns
3. Visualize numerical columns
4. Visualize categorical columns
5. Check the missing value
6. Check the correlation for each variables

>For **Step 3**, it consists of 5 stages:
1. Drop the missing value row
2. Check Outliers
3. Check duplicate data
4. Rerun EDA
5. Encode categorical data - create dummy variables

>For **Step 4**, we'll perform:
1. Split data into training and testing subsets
2. Scaling numerical values using Standard Scale (except dummy variables)
3. Use SMOTE to balance data if the data is imbalanced
4. Feature Selection by using Correlation coefficients, RFECV, and VIFs:
 - 4a. Detect multicollinearity by using Correlation Coefficients and Set Correlation thresholds

- 4b. Select features by using Recursive feature elimination with cross-validation (RFECV)
- 4c. Reduce features further based on the correlation coefficients from step 4a and VIFs, including Check VIF of all features and Drop feature based on correlation coefficients and VIFs (need all VIFs<10). Then, Set up a final train and test datasets with the best features

>For **Step 5**,we'll perform Logistic Regression.
The procedures of this step include:
1. Model creation
2. Test the model to predict using the test dataset
3. Use Statsmodels summary to get the final model's R^2, P value, intercept, and coefficients.
4. Use the final Model for Prediction

Our data is from the IBM company which means it is very useful and realistic in our real business companies, and every company has employee attrition problems. Our analysis could help the company to predict whether the employee would leave the company in the next few years, so the company could take some actions to stay the employees in advance.

After we download this excel, we upload it to the Github for store and for Jupyter use. In the process of statistical analysis, we will check the percentage of attrition in each independent variable. Then, we plan to go through the distributions for numerical attributes and histograms for categorical variables. In addition, we can know the maximum, minimum, mean numbers of each numerical variable. We may use the domain knowledge to judge if there is any unmeaningful attribute or tell us if there are outliers that should be deleted. Other than that, we may utilize the mean or mode to impute the missing values if they are less than 5%. The correlations between one and other variables also are an important part of statistical analysis. We must check the correlations for each variable. It is beneficial for us to discover multicollinearity and do the feature selections.

For the graphics, we plan to use histograms and box plots to show the distribution and outliers for the EDA part. Then use the heatmap to show the coefficient relationship between all variables. We may also use bar plots to show the relationship between some explanatory variables and response variables. Finally, we may use some other graphs to explain our model, like Area Under the Curve(AUC).

When we build the models, we are going to select at least three ways to do that, such as KNN, SVM, Decision tree, and Random forest. To compare the performance of three different models, we will use P-value, R scores, metrics(Accuracy, precision, recall), Root Mean Square Error(RMSE) to define the best model. We will select the best model depending on the main metric. Different purposes use different metrics.

Other than that, we do not have a plan for combining the output of your three individual models into a single ensemble model. At present, since we didn't learn much about the ensemble models, we will explain how to use it later after we learn it.

Last but not least, considering this Data Science project is really helpful to strengthen our capabilities. All of our members would work together for every step, which could make sure we have the best performance. More importantly, we all could have a full understanding for the whole project.