Employee Turnover Behavior Analysis

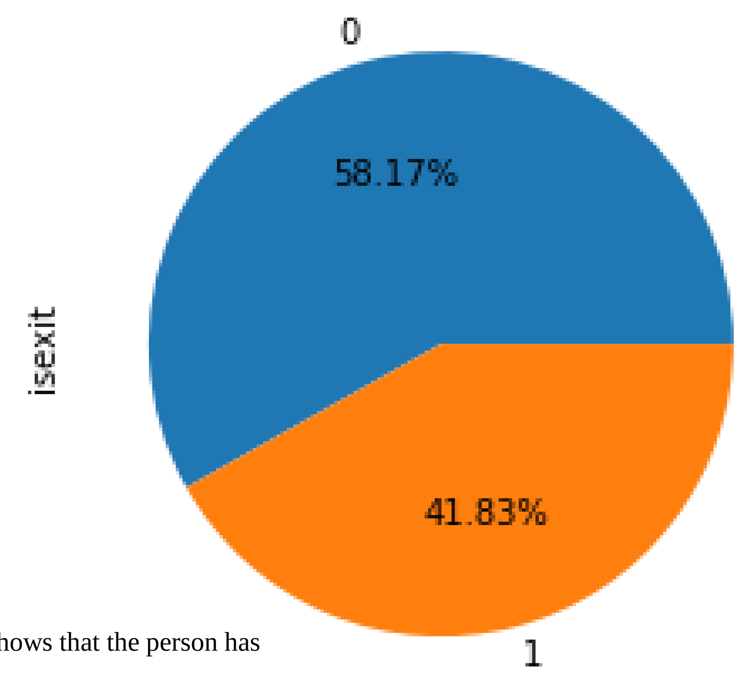
Yujia Yang, Shi Lu, Li Zhu
University of Michigan School of Information

Background

Employee turnover has been a long-standing issue for overall performance and human resource management for organizations. Discovering patterns for turnover and identifying factors that influence turnover behavior are crucial tasks which would help humane resource personales to make better hiring decisions and to improve working environment for the employees. The current project aims to identify patterns and factors that could effectively predict the possible turnover of employees.

Dataset

287229 entries of employment information including the start and end time of a job, if the employee quit the job, industry type of the company, position level, month of working experience and etc[1].



41.83% of the data entries shows that the person has left the job

Previous Work

Zhu et al. [1]who publish the dataset trained machine learning models to predict future turnover possibility using selected features in the dataset. However, based on the feature description of the dataset, we found possible **data leakage** for such method.

Feature "end_time"(calendar year that the employment ends), is an indication of the fact that the employee has left the job. When feature "end_time" is before 2019, feature "isexist" (boolean, if the employment ended) would be "1". Therefore, it is meaningless to train models to predict turnover behavior using entries for employments that ends before 2019. Also, features related to "end_year", such as "GDP"(the GDP at the "end_year", will be source of data leakage.

Reference

[1]Zhu, Qianwen, 2019, "Replication Data for: CoxRF: An employee turnover prediction method based on survival analysis", https://doi.org/10.18170/DVN/NHUSYE, Peking University Open Research Data Platform, V1

Classifier to Predict if the Employee Left the Job

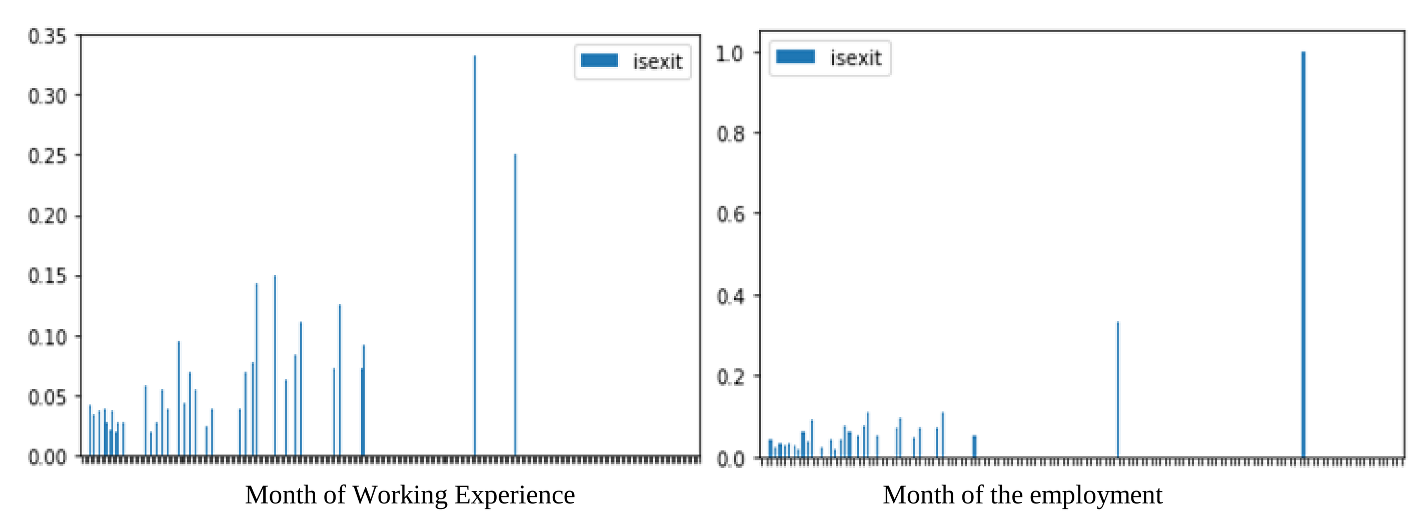
Data: entries with feature "end_year" in 2019 (employees stay or left the job at 2019), N = 3095

Random Forest Classifier

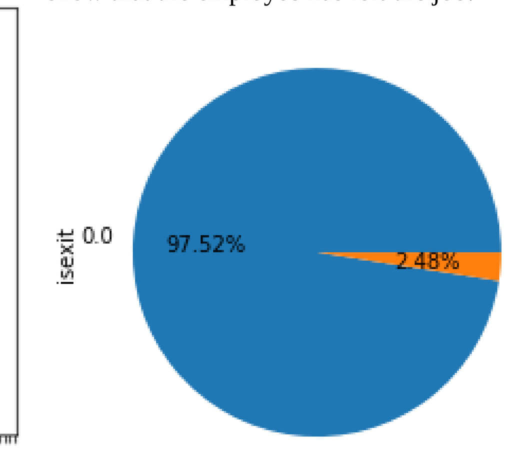
Features: gender, company id, end year of the job, GDP, industry type, position level, number of turnover times, highest degree, length of the employment("timelength"), school type of the highest degree, month of working experience("has_timelength")

Label: If the employee left the job or not

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       751
           1       1.00      0.95      0.97        20

    accuracy                           1.00       771
   macro avg       1.00      0.97      0.99       771
weighted avg       1.00      1.00      1.00       771

test auc: 0.9987029831387808
```



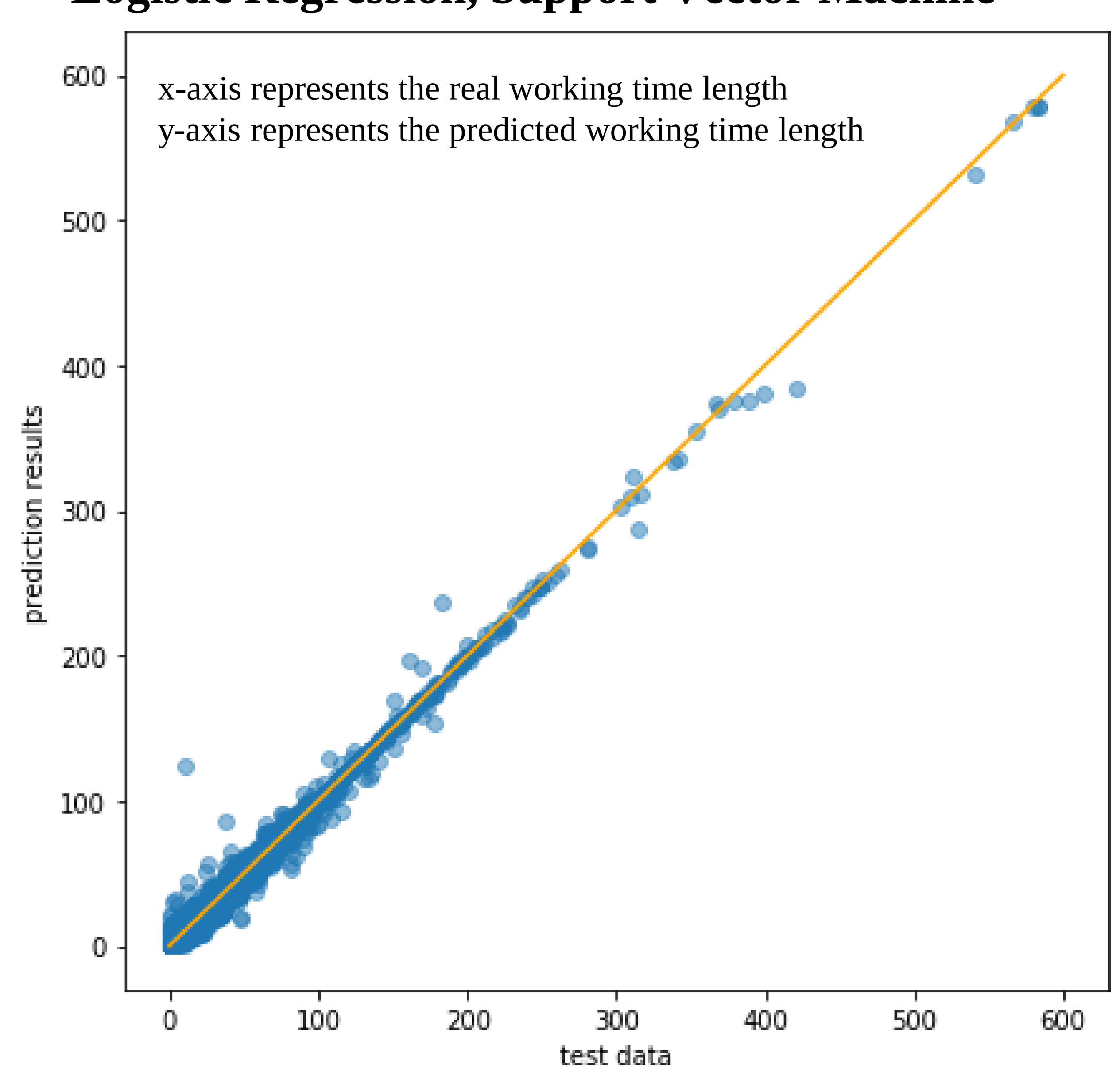
For 2019 employments, 2.48% of the entries show that the employee has left the job.



Regression to Predict the Employee's Total Month of Working Experience

Data: Entries for those who have left the job

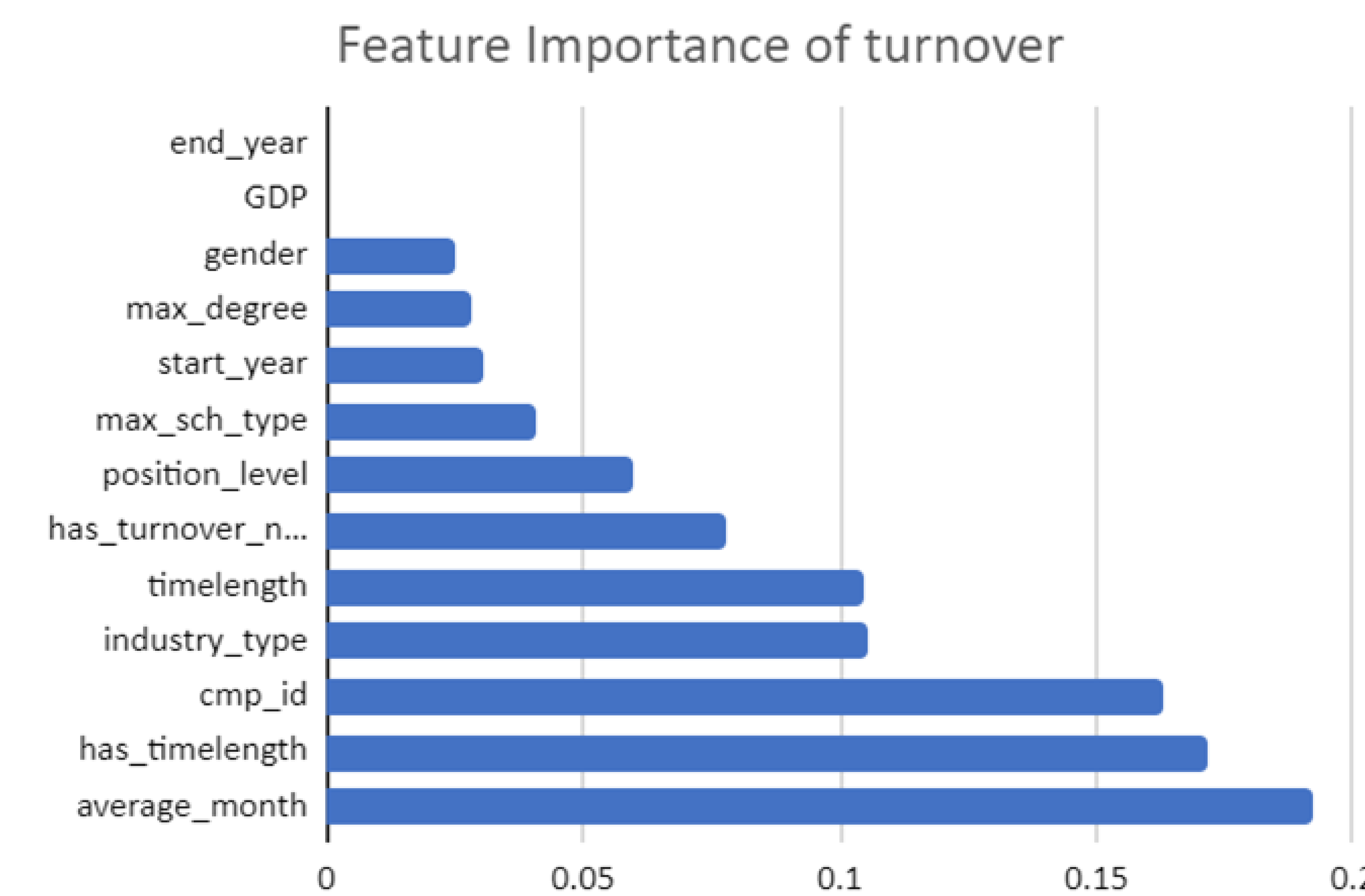
Logistic Regression, Support Vector Machine

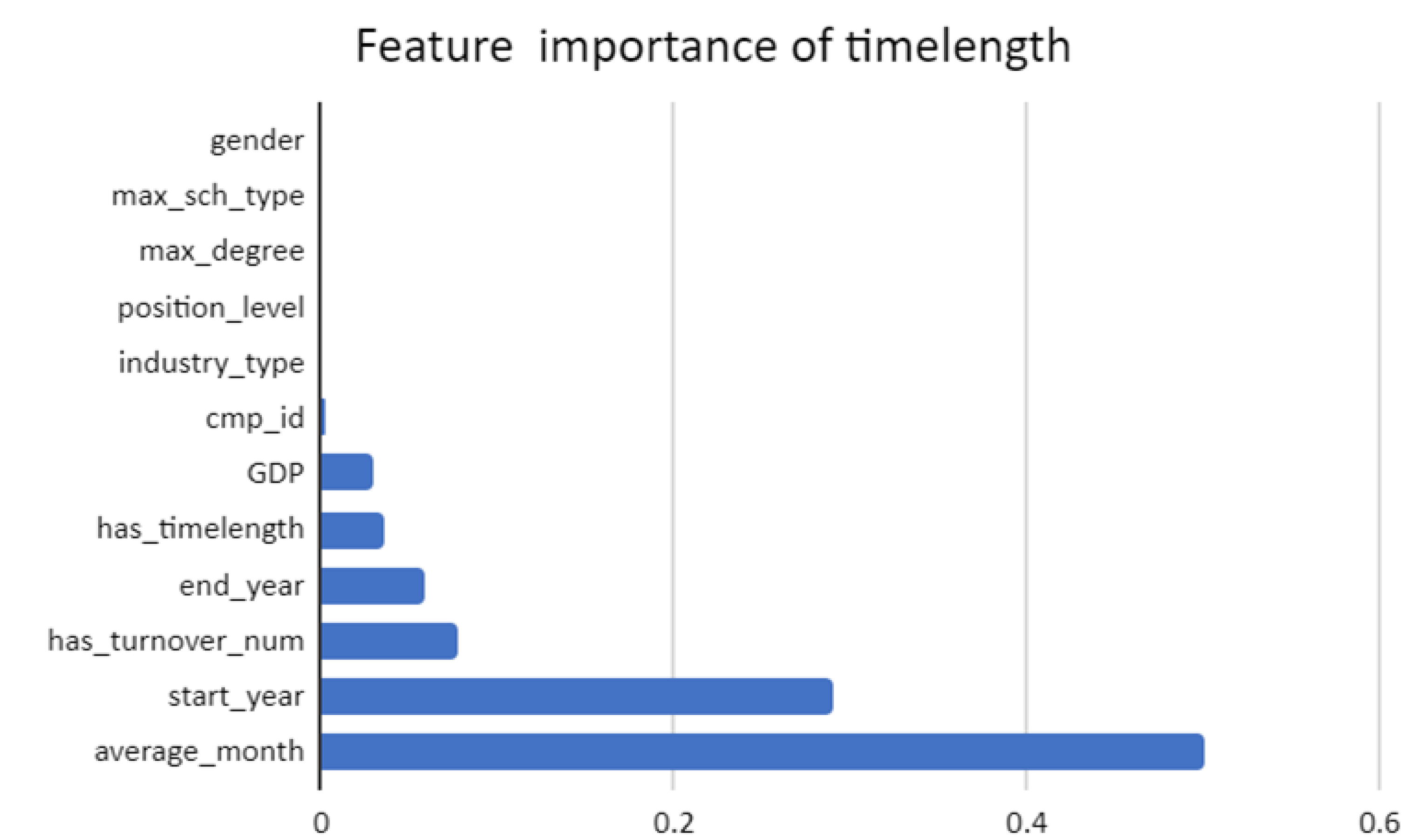


Features:
gender, company id,
end year of the job, GDP,
industry type,
position level,
number of turnover times,
highest degree,
month of the employment (timelength),
school type of the highest degree

Labels:
Month of working experience(has_timelength)

Key Finding





Discussion

For the final features importance, we find that average working time length per job ranked No.1 in both classification and regression model. It means the main factor for an employee to make a job change decision is not because of the company's environment or the current economic environment, but rather the personality of the person.

The current project focus on selecting suitable variables for prediction rather than experimenting with models to optimizing performance, which could be a task in the future work. In this dataset ,we have a lot of website user data but we didn't use them. However, people start to be active in linkedin when they want to find a new job. In this project we didn't use those data from social application but they may be good features to use.