Exploration and Modeling of Attrition Data

Due on May 8, 2020
Knowledge Discovery & Data Mining
CS513B−Spring 2020
Professor Khasha Dehnad

Daniel Kadyrov
10455680

1 Introduction

An attrition dataset was used to predict if an employee is active or terminated based features like age, sex, and education levels. Modeling was performed using a Random Forest classifier and a Support Vector Machine.

2 Data Preprocessing

Initial data features columns including annual and hourly rates, ethnicity, age, sex, job group, first job, and education level. Columns like employee id, termination year, job code, and referral source were removed because they had missing data or unnecessary data for the classification. Status, whether the individual is active or terminated, was selected as the target column and factorized.

Table 1: Data before Processing

EMP_ID	ANNUAL_RATE	HRLY_RATE	JOBCODE	...	STATUS	JOB_GROUP	...
3285941608	33615	22	71850	...	T	Support	...
3687079832	70675	40	59806	...	A	Support	...
7209970080	34320	23	60311	...	A	Support	...
9084013977	103199	59	16233	...	T	Finance	...
4566148978	141801	71	64415	...	A	Marketing	...

2.1 Factorization

Features with a wide range or categorical data needed to be factorized. Annual rate was split based on \$20,000, \$50,000, \$75,000, \$100,000, and \$2,000,000. Hourly rate was split based on \$25, \$50, \$75, \$100, and \$1000. Age was split based on 20, 30, 40, 50, 60, 100. Hire month was split into quarters of a year, Q1, Q2, Q3, and Q4. Ethnicity, sex, marital status, number of teams, first job, travel requirements, disabled, veteran, job group, and education were factorized.

Table 2: Data after Processing

annual rate	hourly rate	ethnicity	sex	marital	satisfaction	...	education	...
2	1	0	0	0	4	...	0	...
3	2	1	1	1	3	...	1	...
2	1	2	0	1	5	...	1	...
5	3	1	0	1	2	...	1	...
5	3	1	0	1	4	...	1	...

2.2 Exploration of Data

There were 21 total number of features with 9612 rows of data. Examining the correlations between the different features with their affect on the status of the employee.

Table 3: Feature Correlation

Feature	Correlation
annual rate	0.178426
hourly rate	0.177944
ethnicity	-0.004023
sex	0.015205
marital	0.014455
satisfaction	0.015613
number of teams	-0.014027
hire month	0.001065
first job	0.005801
travel	-0.003475
rating	-0.001225
disabled emp	-0.008042
disabled vet	-0.003469
education	-0.018809
group	0.032679
prevyr_1	0.148430
prevyr_2	0.163136
prevyr_3	0.177668
prevyr_4	0.213362
prevyr_5	0.220317

3 Modeling

3.1 Feature Selection

Feature selection was performed using Recursive Feature Elimination with a logistic regression model. The following 10 features were selected by the algorithm.

Table 4: Feature Selection

annual rate	hourly rate	sex	marital	age	disabled emp	group	prevyr_1	prevyr_4	prevyr_5
2	1	0	0	3	0	0	0	0	0
3	2	1	1	1	0	0	3	2	3
2	1	0	1	1	0	0	3	2	3
5	3	0	1	3	0	1	0	0	0
5	3	0	1	3	0	2	2	2	2

3.2 Training Test Split

The data was split was into 70% training and 30% test subsects.

3.3 Random Forest

Random Forest classification was performed on the training data and the mdoel was used to predict the test data for accuracy comparison. The classificiation report for the Random Forest model:

Table 5: RF Classification Report

	precision	recall	f1-score	support
T	0.571429	0.502749	0.534893	1273.000000
A	0.641156	0.702048	0.670222	1611.000000
accuracy	0.614078	0.614078	0.614078	0.614078
macro avg	0.606293	0.602399	0.602558	2884.000000
weighted avg	0.610379	0.614078	0.610488	2884.000000

3.4 Support Vector Machine

Support Vector Machine classification was performed on the training data and the model was used to predict the test data for accuracy comparison. The classification report for the SVM model:

Table 6: SVM Classification Report

	precision	recall	f1-score	support
T	0.622665	0.392773	0.481696	1273.000000
A	0.628544	0.811918	0.708559	1611.000000
accuracy	0.626907	0.626907	0.626907	0.626907
macro avg	0.625604	0.602346	0.595127	2884.000000
weighted avg	0.625949	0.626907	0.608421	2884.000000

3.5 Model Accuracy

The accuracy score of the models was performed based on their predictions of the test data.

Table 7: Model Accuracy

rf	svc
0.614078	0.626907

4 Conclusion

Random Forest classification and Support Vector Machine were used to predict the attrition of employees based on features in a dataset. The SVM performed better than the RF classifier.