Agenda

- Data Preprocessing and EDA (Exploratory Data Analysis)
- Feature Engineering (One-hot encoding & PCA)
- Classification Analysis
  - Naive Bayes
  - KNN
  - Support Vector Machine (SVM)
  - Random Forest
  - XGBoost
- Evaluation and Conclusion

Data preprocessing & EDA

Data preprocessing

- Deleted the meaningless attributes:
  a. *There are 27 variables and 9612 obs. in this data. We deleted 2 variables (EMP_ID and JOBCODE) which are meaningless.*
  b. *TERMINATION_YEAR is determined by STATUS, so remove it.*

- Deal with the missing value:
  a. *Deleted the missing value in ETHNICITY.*
  b. *Converted the missing value in REFERRAL SOURCE change to "Unknown" category.*

Data preprocessing

- Regrouping JOB_GROUP:

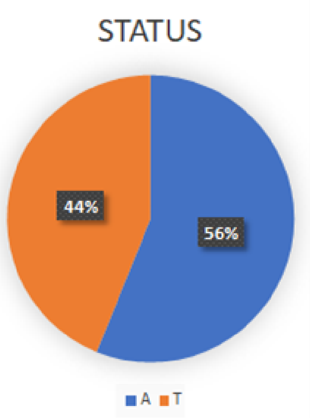
  There are too many levels in JOB_GROUP, therefore, we regroup them to 6 levels based on the job title meaning.

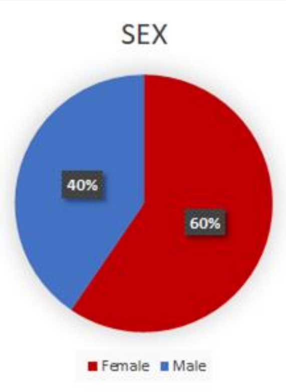
- Normalize the numeric variables by Min-max normalization.

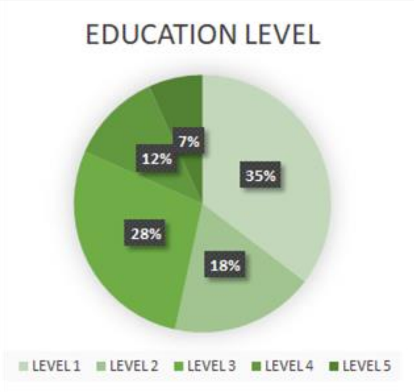
- Factorize the categorical variables.

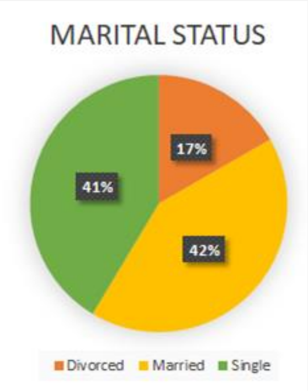
- Convert Outlier to mean: $Leverage = h = \frac{1}{N} + \frac{(X - X)^2}{\sum(X - X)^2}, \ Outlier\ in\ X\ if\ h > 4/N$

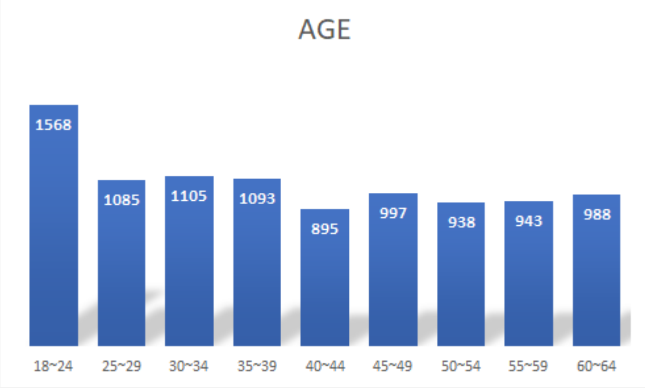
Basic description

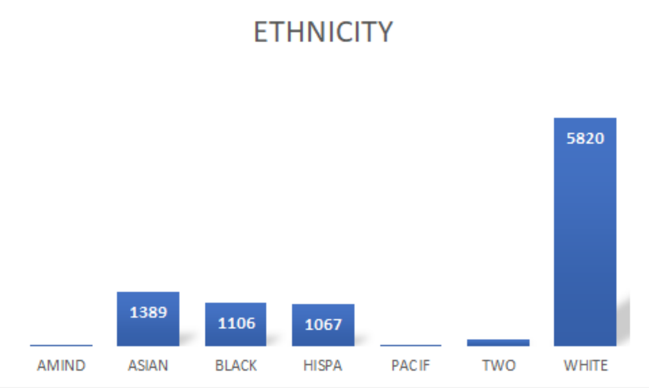








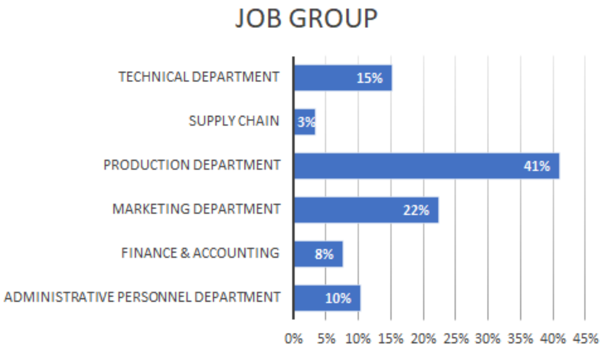


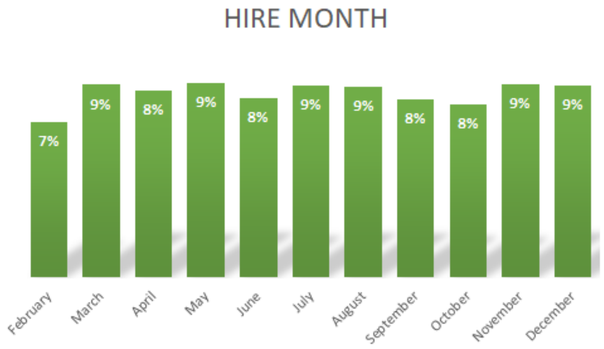


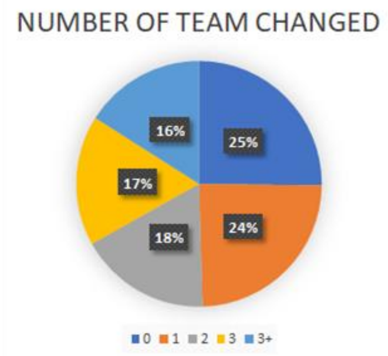
- 44% of employees are terminal and 56% of employees are still in the company
- The main employee might be female, education level lower and white

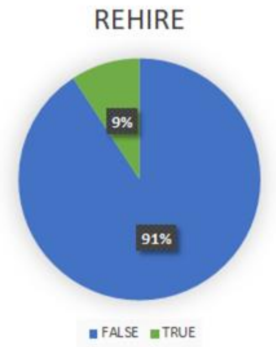
Job description

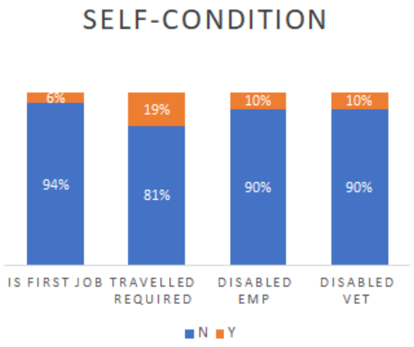
- 41% of employees work in the production department
- About 50% of employees change team less than 1 time
- Only 9% of employees are rehired
- Most employees do not work for the first time
- The rate of DISABLE EMP and VET are both around 10%


JOB GROUP


HIRE MONTH

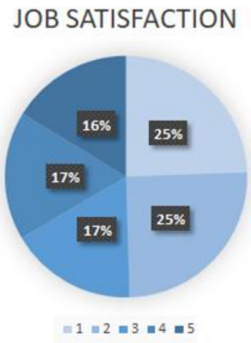

NUMBER OF TEAM CHANGED

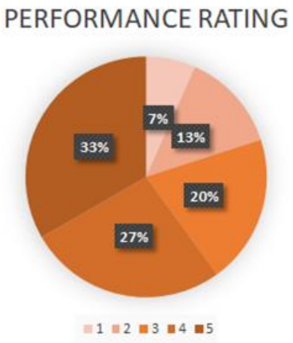

REHIRE

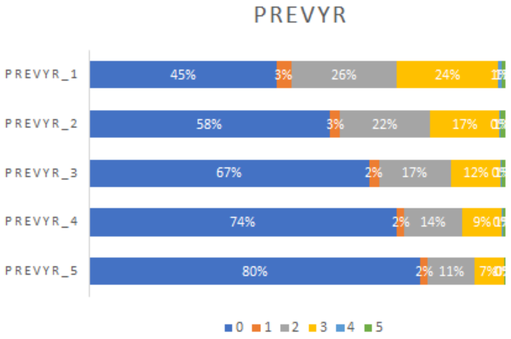

SELF-CONDITION

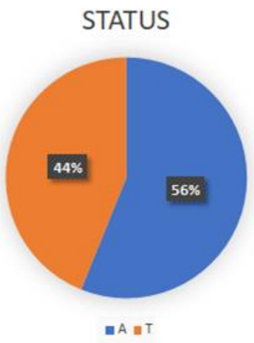
Rating score and Status situation

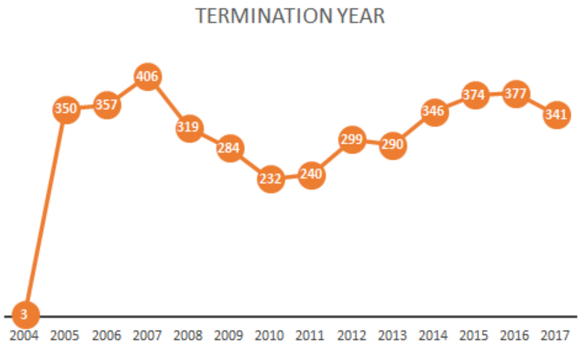
- 51% of employees feel satisfy their job (more than 3)
- 80% of employees are rated more than 3 point
- 0 of the PREVYR1~5 are more than 45%
- The highest termination year is 2007, and the high termination year are 2005~2007 and 2014~2017


JOB SATISFACTION


PERFORMANCE RATING

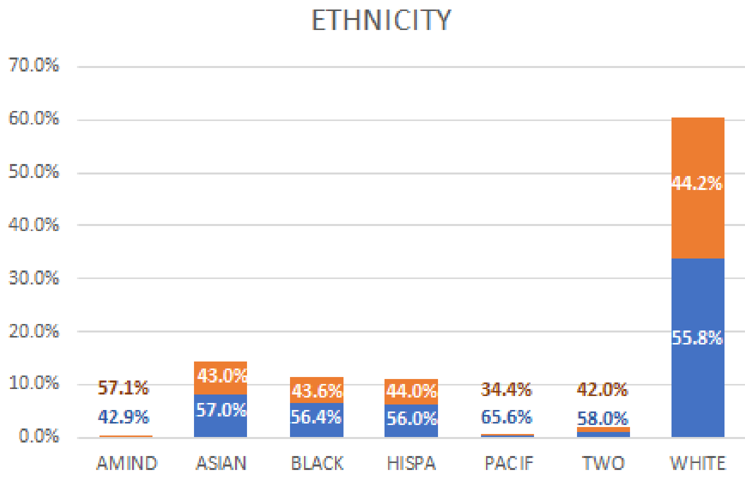

PREVYR

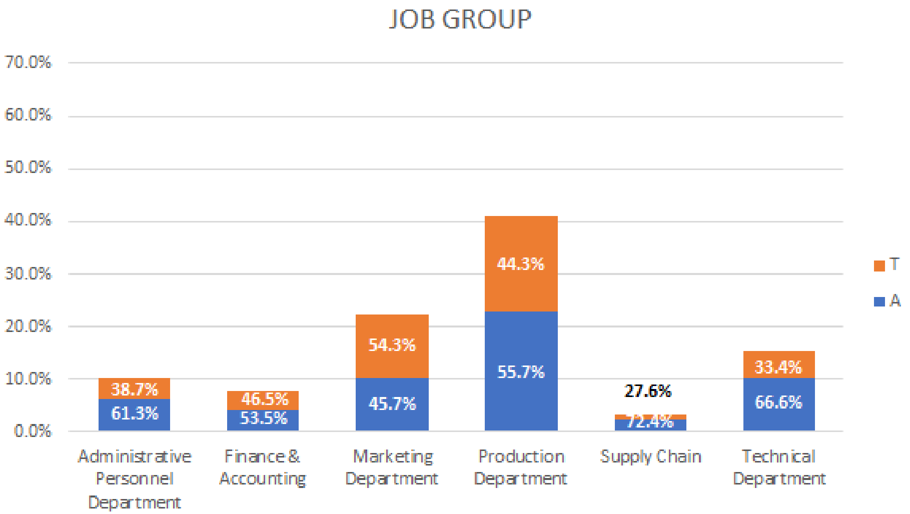

STATUS


TERMINATION YEAR

Status v.s. Ethnicity & Job Group

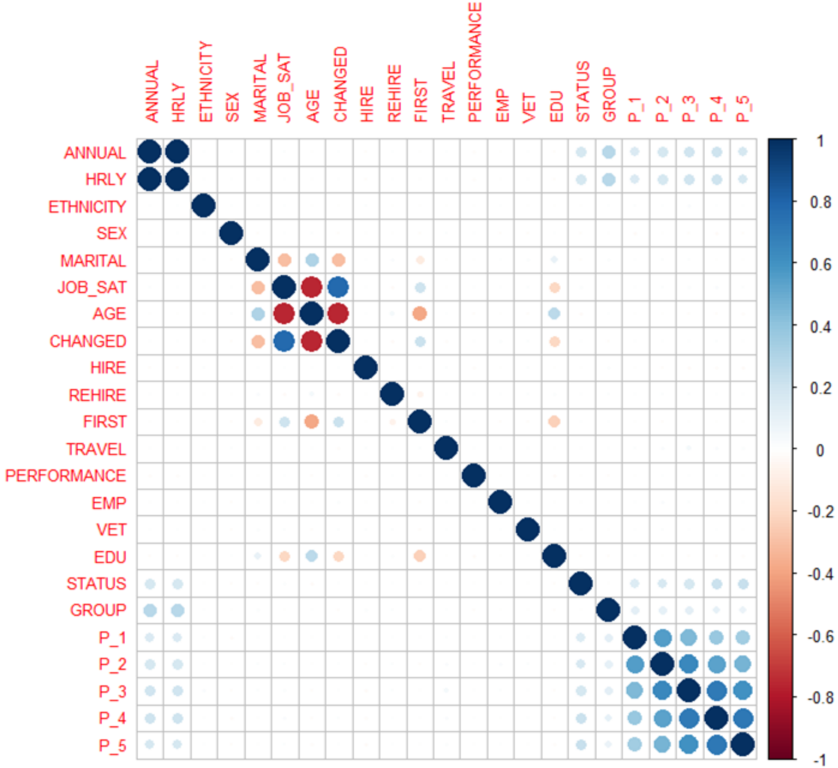
- Most ethnicity have no significant difference in STATUS, however, the rate of AMIND terminal is more than other ethnicity. (more than 44%)
- The terminal rate of Marketing Department is more than other department. Otherwise, the rate of Administrative Personnel, Technical and Supply Chain these 3 departments are less than 44%


ETHNICITY


JOB GROUP

Variable correlation

- From the variable correlation picture, we found out some variables are positive relative: MARITAL & AGE, EDU & AGE, JOB_SAT & TEAM CHANGED, JOB_SAT & FIRST JOB, GROUP & many rate scores, PREVYRs
- Negative relative: JOB_SAT & AGE, JOB_SAT & TEAM CHANGED, AGE & FIRST JOB, EDU & FIRST JOB



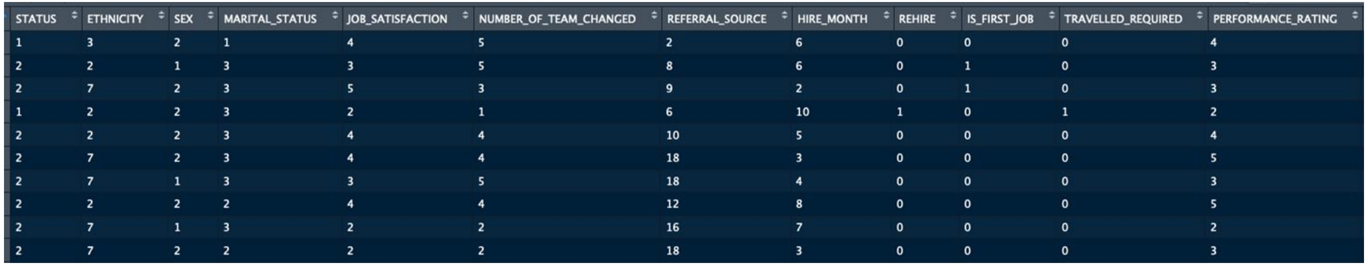
Feature Engineering
&
Classification Analysis

Road Map

1. Select influential variables using Feature Engineering (One-hot Encoding & PCA) .
2. Improve performance using ensemble methods (Random Forest & XGBoost).
3. Optimize models using Cross Validation.
4. Optimize parameters using Grid Search.

Feature Engineering

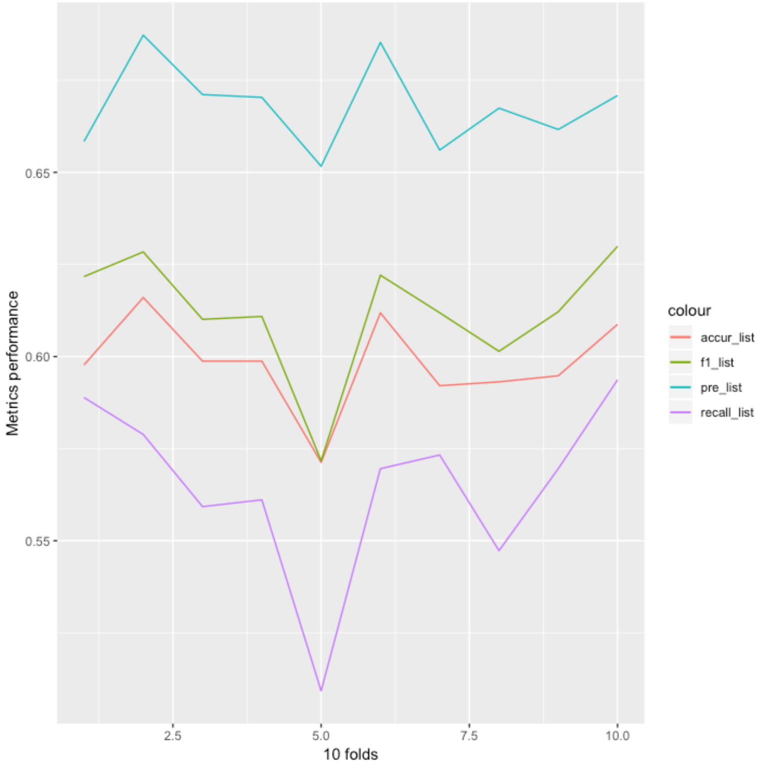
1. Convert multi-level attributes to dummy variables using one-hot encoding
2. Select most relevant variable using PCA (threshold = 80%)



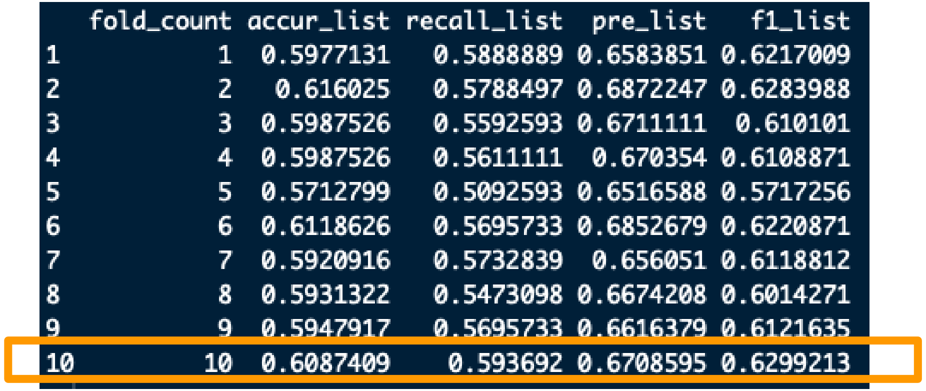
STATUS	ETHNICITY	SEX	MARITAL_STATUS	JOB_SATISFACTION	NUMBER_OF_TEAM_CHANGED	REFERRAL_SOURCE	HIRE_MONTH	REHIRE	IS_FIRST_JOB	TRAVELLED_REQUIRED	PERFORMANCE_RATING
1	3	2	1	4	5	2	6	0	0	0	4
2	2	1	3	3	5	8	6	0	1	0	3
2	7	2	3	5	3	9	2	0	1	0	3
1	2	2	3	2	1	6	10	1	0	1	2
2	2	2	3	4	4	10	5	0	0	0	4
2	7	2	3	4	4	18	3	0	0	0	5
2	7	1	3	3	5	18	4	0	0	0	3
2	2	2	2	4	4	12	8	0	0	0	5
2	7	1	3	2	2	16	7	0	0	0	2
2	7	2	2	2	2	18	3	0	0	0	3

Naive Bayes

Metrics Performance of 10 folds



Detail Results of 10 folds



```
   fold_count accur_list recall_list  pre_list   f1_list
1           1  0.5977131   0.5888889 0.6583851 0.6217009
2           2   0.616025   0.5788497 0.6872247 0.6283988
3           3  0.5987526   0.5592593 0.6711111  0.610101
4           4  0.5987526   0.5611111  0.670354 0.6108871
5           5  0.5712799   0.5092593 0.6516588 0.5717256
6           6  0.6118626   0.5695733 0.6852679 0.6220871
7           7  0.5920916   0.5732839  0.656051 0.6118812
8           8  0.5931322   0.5473098 0.6674208 0.6014271
9           9  0.5947917   0.5695733 0.6616379 0.6121635
10         10  0.6087409    0.593692 0.6708595 0.6299213
```

Mean Results of 10 folds:

Accuracy: 0.598

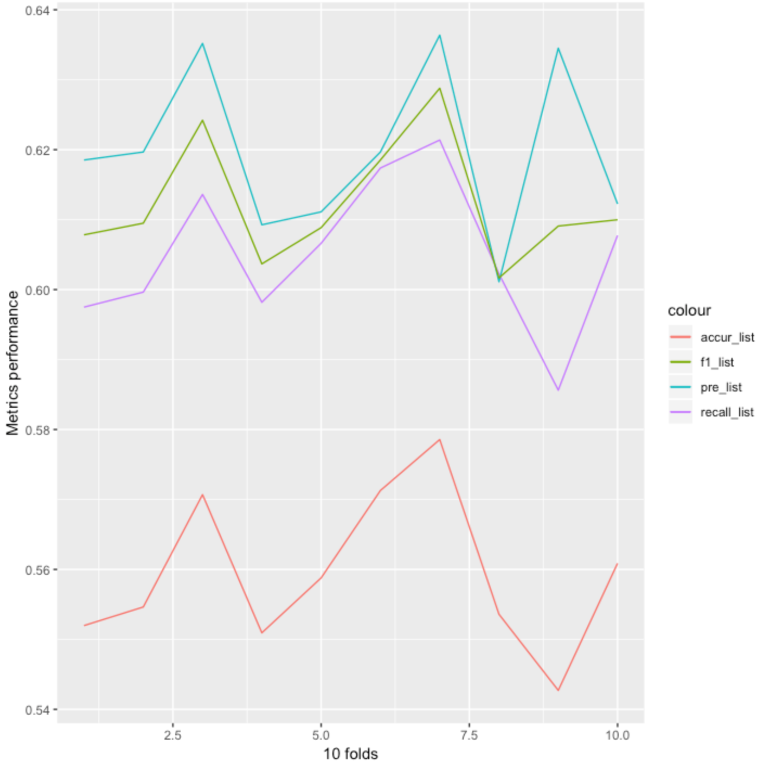
Recall: 0.565

Precision: 0.668

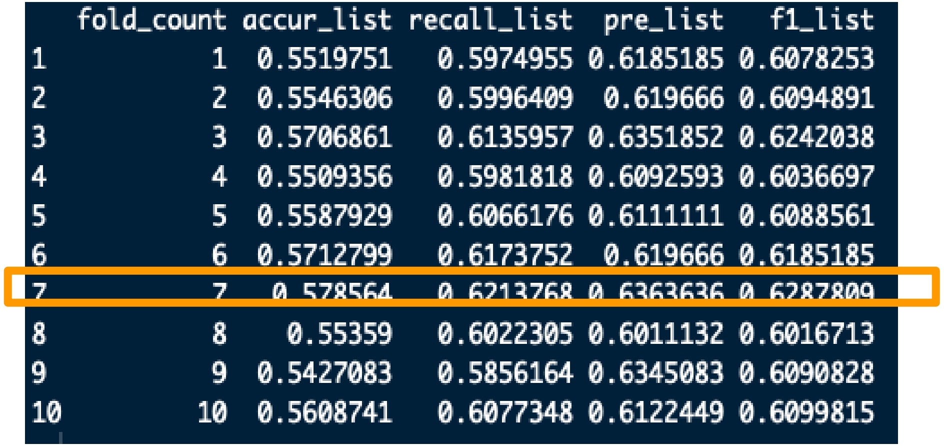
F1_Score: 0.612

KNN

Metrics Performance of 10 folds



Detail Results of 10 folds



```
   fold_count accur_list recall_list  pre_list   f1_list
1           1  0.5519751   0.5974955 0.6185185 0.6078253
2           2  0.5546306   0.5996409  0.619666 0.6094891
3           3  0.5706861   0.6135957 0.6351852 0.6242038
4           4  0.5509356   0.5981818 0.6092593 0.6036697
5           5  0.5587929   0.6066176 0.6111111 0.6088561
6           6  0.5712799   0.6173752  0.619666 0.6185185
7           7   0.578564   0.6213768 0.6363636 0.6287809
8           8    0.55359   0.6022305 0.6011132 0.6016713
9           9  0.5427083   0.5856164 0.6345083 0.6090828
10         10  0.5608741   0.6077348 0.6122449 0.6099815
```

Mean Results of 10 folds:

Accuracy: 0.56

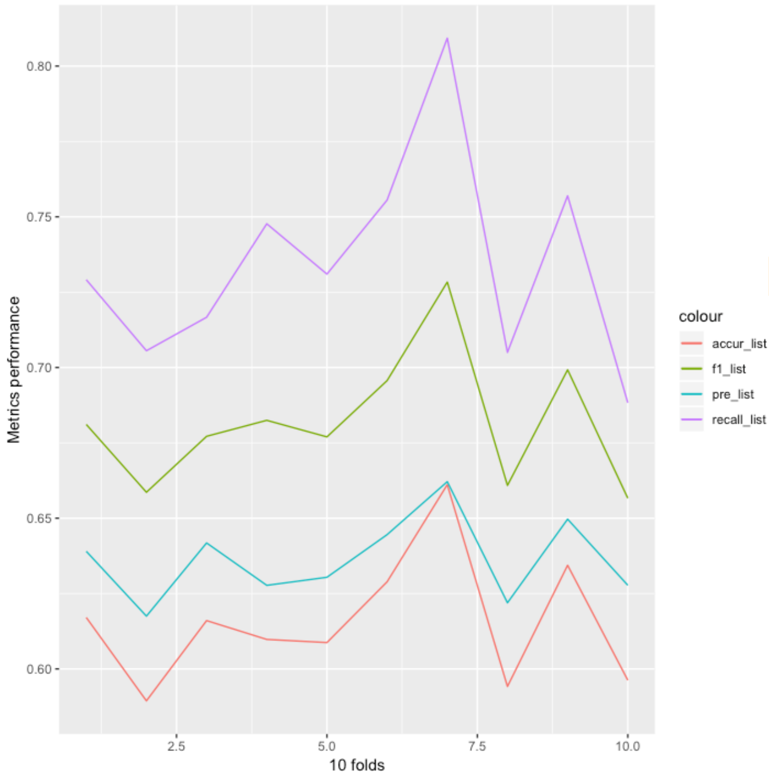
Recall: 0.605

Precision: 0.620

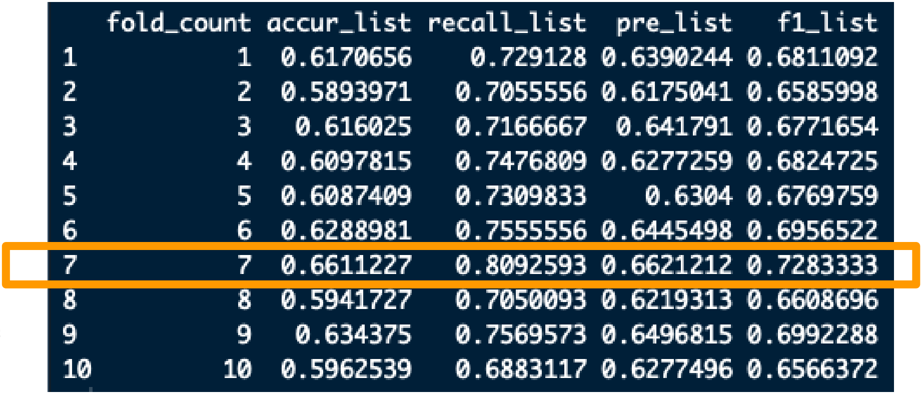
F1_Score: 0.612

Support Vector Machine (SVM)

Metrics Performance of 10 folds



Detail Results of 10 folds



```
   fold_count accur_list recall_list  pre_list   f1_list
1           1  0.6170656    0.729128 0.6390244 0.6811092
2           2  0.5893971   0.7055556 0.6175041 0.6585998
3           3   0.616025   0.7166667  0.641791 0.6771654
4           4  0.6097815   0.7476809 0.6277259 0.6824725
5           5  0.6087409   0.7309833    0.6304 0.6769759
6           6  0.6288981   0.7555556 0.6445498 0.6956522
7           7  0.6611227   0.8092593 0.6621212 0.7283333
8           8  0.5941727   0.7050093 0.6219313 0.6608696
9           9   0.634375   0.7569573 0.6496815 0.6992288
10         10  0.5962539   0.6883117 0.6277496 0.6566372
```

Mean Results of 10 folds:

Accuracy: 0.616

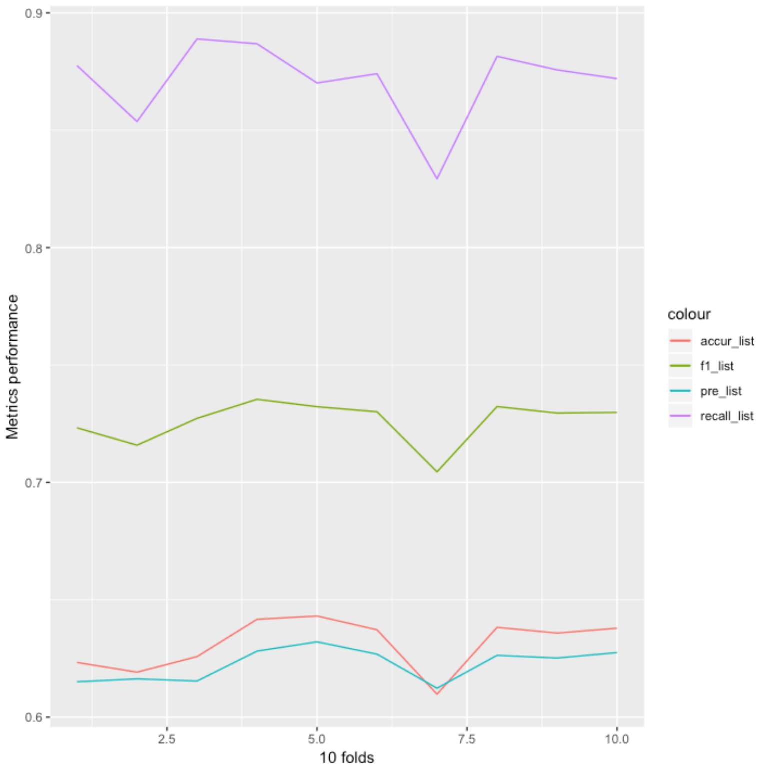
Recall: 0.734

Precision: 0.636

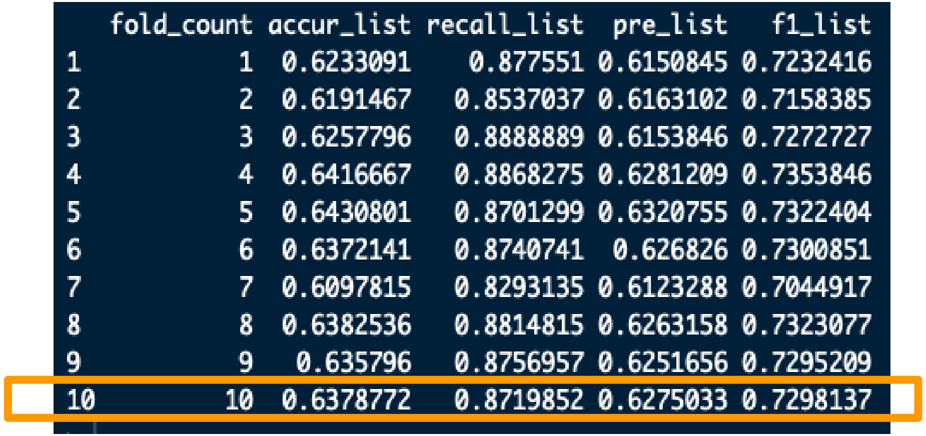
F1_Score: 0.682

Random Forest

Metrics Performance of 10 folds



Detail Results of 10 folds



```
   fold_count accur_list recall_list  pre_list   f1_list
1           1  0.6233091    0.877551 0.6150845 0.7232416
2           2  0.6191467   0.8537037 0.6163102 0.7158385
3           3  0.6257796   0.8888889 0.6153846 0.7272727
4           4  0.6416667   0.8868275 0.6281209 0.7353846
5           5  0.6430801   0.8701299 0.6320755 0.7322404
6           6  0.6372141   0.8740741  0.626826 0.7300851
7           7  0.6097815   0.8293135 0.6123288 0.7044917
8           8  0.6382536   0.8814815 0.6263158 0.7323077
9           9   0.635796   0.8756957 0.6251656 0.7295209
10         10  0.6378772   0.8719852 0.6275033 0.7298137
```

Mean Results of 10 folds:

Accuracy: 0.631

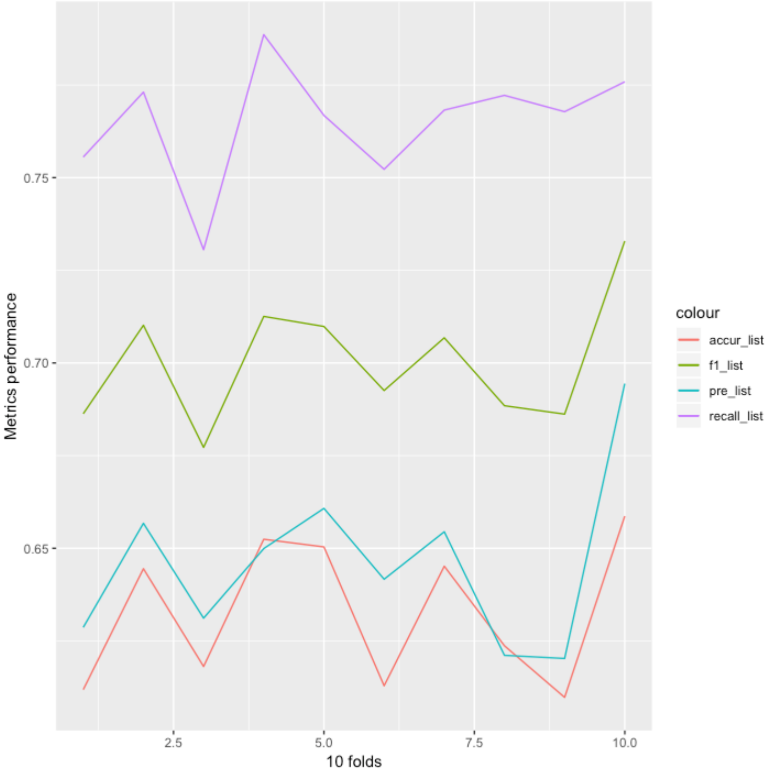
Recall: 0.87

Precision: 0.623

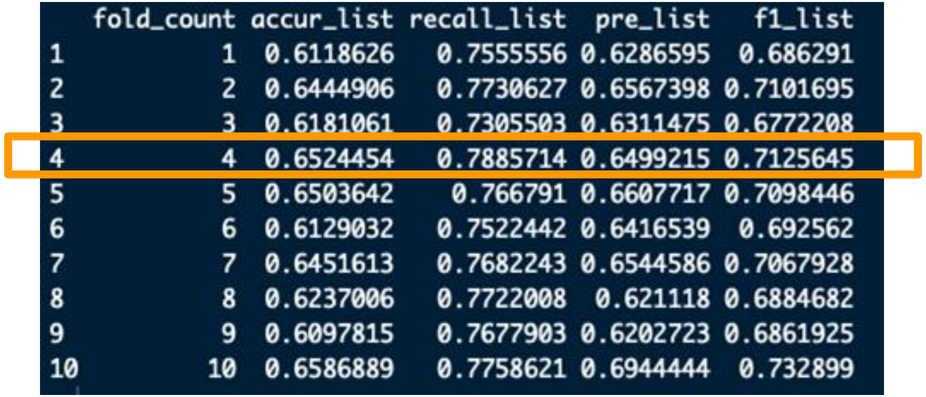
F1_Score: 0.726

XGBoost

Metrics Performance of 10 folds



Detail Results of 10 folds



```
   fold_count accur_list recall_list  pre_list   f1_list
1           1  0.6118626   0.7555556 0.6286595  0.686291
2           2  0.6444906   0.7730627 0.6567398 0.7101695
3           3  0.6181061   0.7305503 0.6311475 0.6772208
4           4  0.6524454   0.7885714 0.6499215 0.7125645
5           5  0.6503642    0.766791 0.6607717 0.7098446
6           6  0.6129032   0.7522442 0.6416539  0.692562
7           7  0.6451613   0.7682243 0.6544586 0.7067928
8           8  0.6237006   0.7722008  0.621118 0.6884682
9           9  0.6097815   0.7677903 0.6202723 0.6861925
10         10  0.6586889   0.7758621 0.6944444  0.732899
```

Mean Results of 10 folds:

Accuracy: 0.633

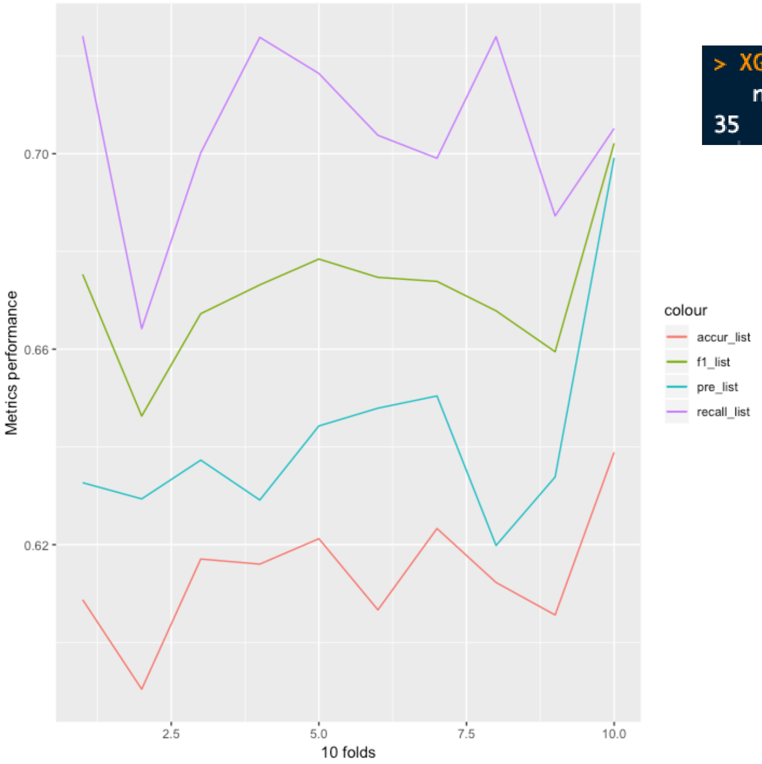
Recall: 0.77

Precision: 0.646

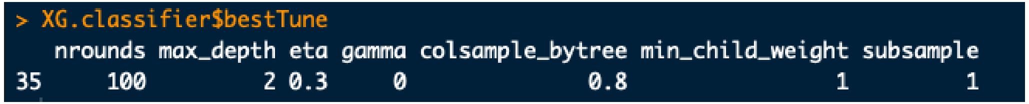
F1_Score: 0.7

XGBoost_Grid Search

Metrics Performance of 10 folds



Best parameters based on Grid Search



```
> XG.classifier$bestTune
   nrounds max_depth eta gamma colsample_bytree min_child_weight subsample
35     100         2 0.3     0              0.8                1         1
```

Mean Results of 10 folds:

Accuracy: 0.614

Recall: 0.704

Precision: 0.642

F1_Score: 0.67

Evaluation and Conclusion

1. One-hot encoding did not work in the case.
2. PCA helped improve accuracy roughly 3%~5%.
3. Among 5 classification algorithms, the two ensemble algorithms are better than the others.
4. Grid search did not work in the case.
5. In the future work, more sophisticated feature engineering need to be implemented. (e.g. resampling, regrouping categorical variables)