Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

Created by:
Muhammad Miftah Thaha
muhammadmiftaht@gmail.com
https://www.linkedin.com/in/miftahthaha/
https://github.com/miftahthaha

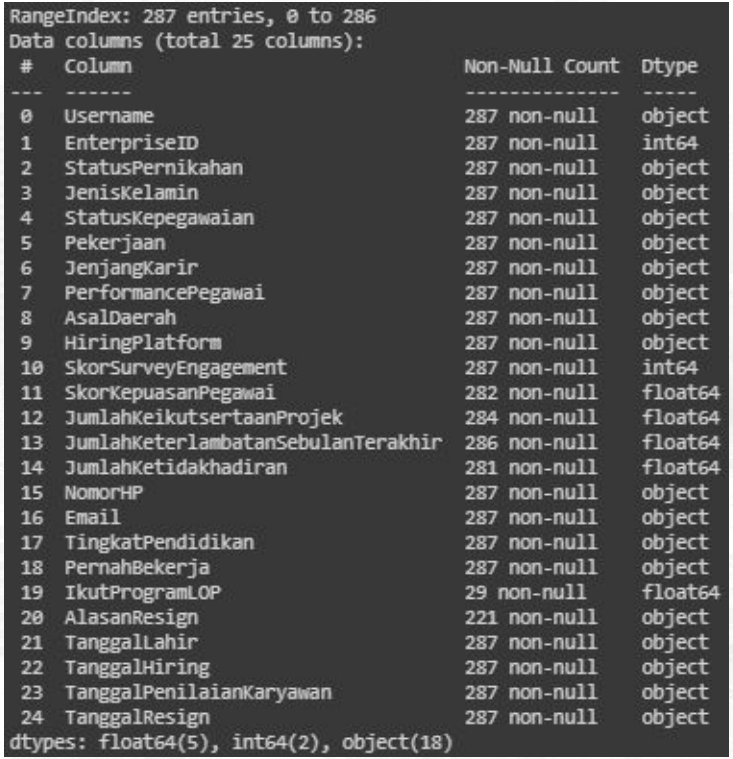
Miftah is a recent graduate with a Bachelor's degree Electrical Engineer who have interest in Data Analytics and Science and have a strong foundation in statistical modeling, data analysis, and programming. As a Junior Data Scientist, he has experience through his final project in building and implementing machine learning models, analyzing complex data sets, and creating visualizations to communicate insights. He is a fast learner with excellent problem-solving skills and a passion for using data to drive business decisions. In addition, he possess strong communication and collaboration skills, having worked on multiple team projects during his studies. With a drive to excel in his field, Miftah is seeking an opportunity to contribute his skills and knowledge to a dynamic and innovative organization as a Junior Data Scientist.

Overview

"Human resources (HR) are the main asset that needs to be managed effectively and efficiently by a company in order to achieve business goals. In this opportunity, we will face a problem related to human resources in a company. Our focus is to find out how to keep employees staying in the current company, which can result in increased costs for recruiting and training new employees. By identifying the main factors that cause employees to feel dissatisfied, the company can immediately address them by creating relevant programs that address employee issues."

Data Preprocessing

- ***Dataset***



```
RangeIndex: 287 entries, 0 to 286
Data columns (total 25 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   Username                          287 non-null    object 
 1   EnterpriseID                      287 non-null    int64  
 2   StatusPernikahan                  287 non-null    object 
 3   JenisKelamin                      287 non-null    object 
 4   StatusKepegawaian                 287 non-null    object 
 5   Pekerjaan                         287 non-null    object 
 6   JenjangKarir                      287 non-null    object 
 7   PerformancePegawai                287 non-null    object 
 8   AsalDaerah                        287 non-null    object 
 9   HiringPlatform                    287 non-null    object 
 10  SkorSurveyEngagement              287 non-null    int64  
 11  SkorKepuasanPegawai               282 non-null    float64
 12  JumlahKeikutsertaanProjek         284 non-null    float64
 13  JumlahKeterlambatanSebulanTerakhir 286 non-null    float64
 14  JumlahKetidakhadiran              281 non-null    float64
 15  NomorHP                           287 non-null    object 
 16  Email                             287 non-null    object 
 17  TingkatPendidikan                 287 non-null    object 
 18  PernahBekerja                     287 non-null    object 
 19  IkutProgramLOP                    29 non-null     float64
 20  AlasanResign                      221 non-null    object 
 21  TanggalLahir                      287 non-null    object 
 22  TanggalHiring                     287 non-null    object 
 23  TanggalPenilaianKaryawan          287 non-null    object 
 24  TanggalResign                     287 non-null    object 
dtypes: float64(5), int64(2), object(18)
```

- Description

  Dataset that contains information related to personal information made by HR Departement.

- Shape

  287 Row and 25 Columns (Feature)

- Datatypes

  Float64 (5 Column), Int64 (2 Column), object (18 Column)

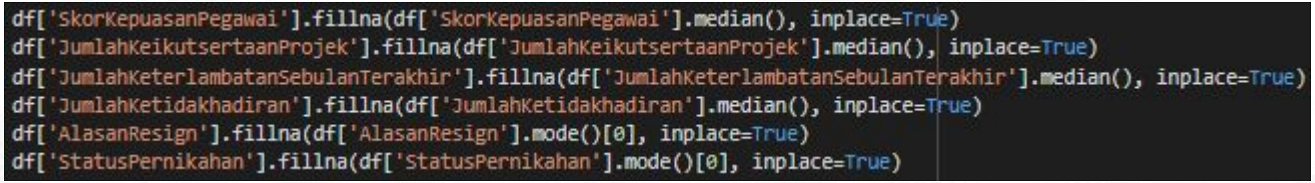
- Missing Values

  Detected in 4 Column

For more details, you can see all file here and code here

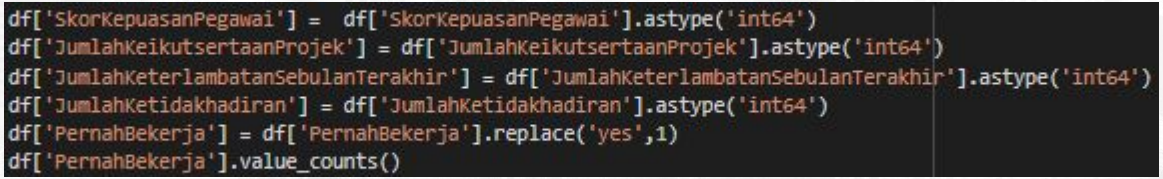
Data Preprocessing

- Null Data



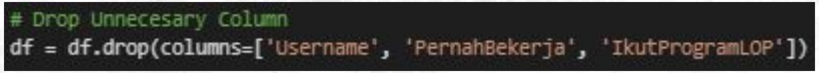
```
df['SkorKepuasanPegawai'].fillna(df['SkorKepuasanPegawai'].median(), inplace=True)
df['JumlahKeikutsertaanProjek'].fillna(df['JumlahKeikutsertaanProjek'].median(), inplace=True)
df['JumlahKeterlambatanSebulanTerakhir'].fillna(df['JumlahKeterlambatanSebulanTerakhir'].median(), inplace=True)
df['JumlahKetidakhadiran'].fillna(df['JumlahKetidakhadiran'].median(), inplace=True)
df['AlasanResign'].fillna(df['AlasanResign'].mode()[0], inplace=True)
df['StatusPernikahan'].fillna(df['StatusPernikahan'].mode()[0], inplace=True)
```

- Adjusting Data Types



```
df['SkorKepuasanPegawai'] =  df['SkorKepuasanPegawai'].astype('int64')
df['JumlahKeikutsertaanProjek'] = df['JumlahKeikutsertaanProjek'].astype('int64')
df['JumlahKeterlambatanSebulanTerakhir'] = df['JumlahKeterlambatanSebulanTerakhir'].astype('int64')
df['JumlahKetidakhadiran'] = df['JumlahKetidakhadiran'].astype('int64')
df['PernahBekerja'] = df['PernahBekerja'].replace('yes',1)
df['PernahBekerja'].value_counts()
```

- Drop Column

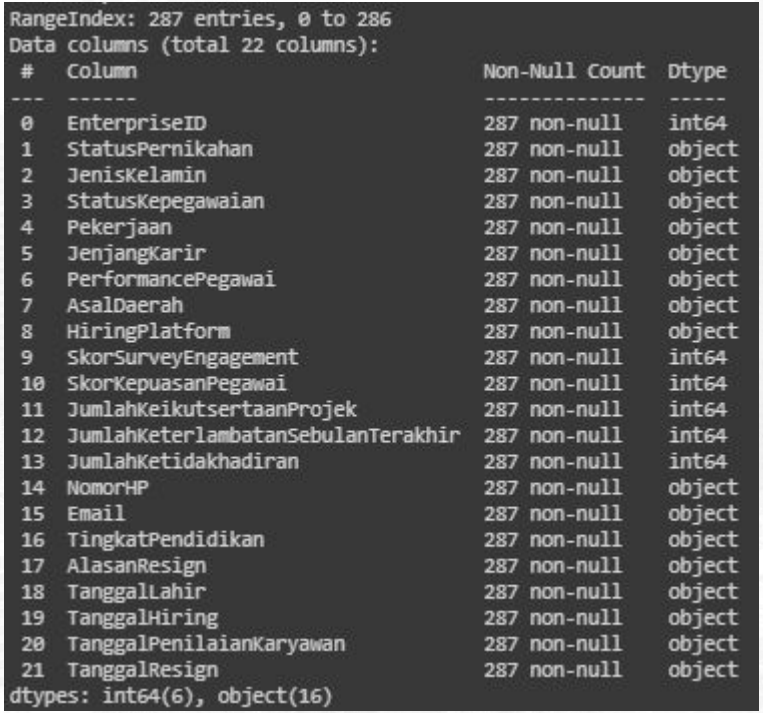


```
# Drop Unnecesary Column
df = df.drop(columns=['Username', 'PernahBekerja', 'IkutProgramLOP'])
```

For more details, you can see all file here and code here

Data Preprocessing

- Dataset after Preprocessing

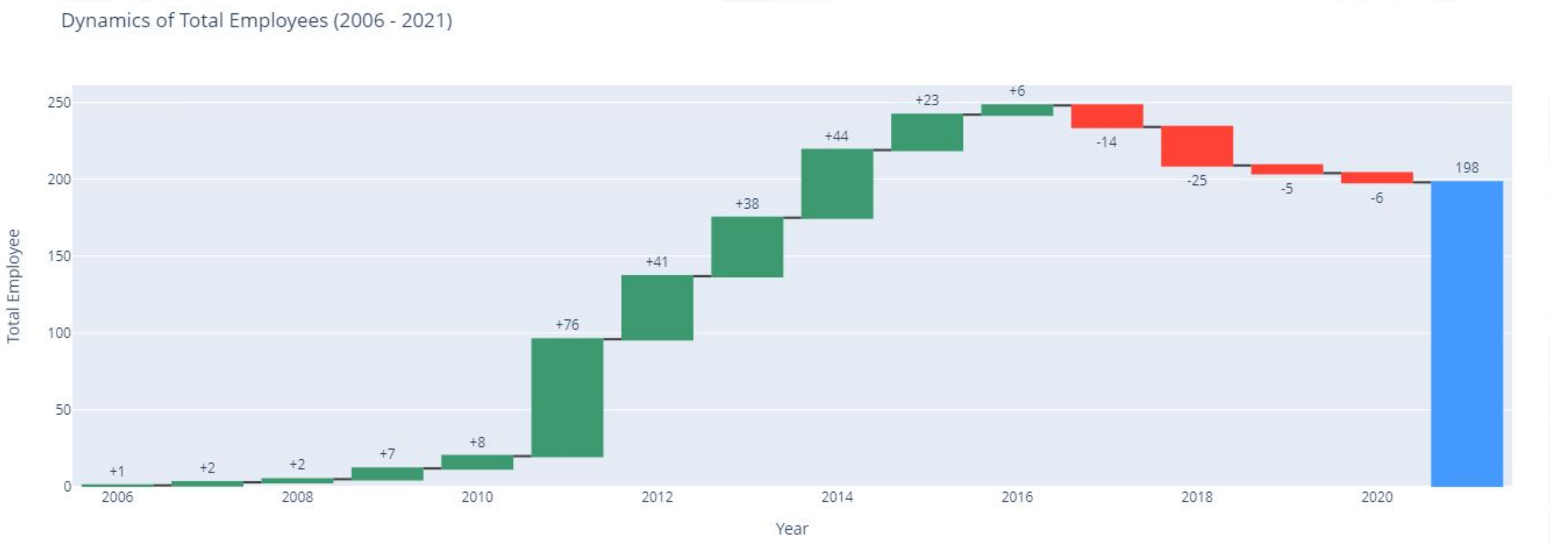


```
RangeIndex: 287 entries, 0 to 286
Data columns (total 22 columns):
 #   Column                              Non-Null Count  Dtype 
---  ------                              --------------  ----- 
 0   EnterpriseID                        287 non-null    int64 
 1   StatusPernikahan                    287 non-null    object
 2   JenisKelamin                        287 non-null    object
 3   StatusKepegawaian                   287 non-null    object
 4   Pekerjaan                           287 non-null    object
 5   JenjangKarir                        287 non-null    object
 6   PerformancePegawai                  287 non-null    object
 7   AsalDaerah                          287 non-null    object
 8   HiringPlatform                      287 non-null    object
 9   SkorSurveyEngagement                287 non-null    int64 
 10  SkorKepuasanPegawai                 287 non-null    int64 
 11  JumlahKeikutsertaanProjek           287 non-null    int64 
 12  JumlahKeterlambatanSebulanTerakhir  287 non-null    int64 
 13  JumlahKetidakhadiran                287 non-null    int64 
 14  NomorHP                             287 non-null    object
 15  Email                               287 non-null    object
 16  TingkatPendidikan                   287 non-null    object
 17  AlasanResign                        287 non-null    object
 18  TanggalLahir                        287 non-null    object
 19  TanggalHiring                       287 non-null    object
 20  TanggalPenilaianKaryawan            287 non-null    object
 21  TanggalResign                       287 non-null    object
dtypes: int64(6), object(16)
```

The dataset comprises of 287 rows and 25 columns, with no apparent duplicate values, although two usernames, boredEggs0 and brainyMagpie7, appear more than once but with distinct values in other features. Due to an issue with the Username feature, it will be removed along with the PernahBekerja feature. The dataset includes null values in some features, with over 80% null values in IkutProgramLOP resulting in its exclusion, while other features such as SkorKepuasanPegawai, JumlahKeikutsertaanProjek, JumlahKeterlambatanSebulanTerakhir, JumlahKetidakhadiran, and AlasanResign have null values below 25% and will require imputation. Furthermore, certain features will need to be transformed from float64 to int64 data type.

For more details, you can see all file here and code here

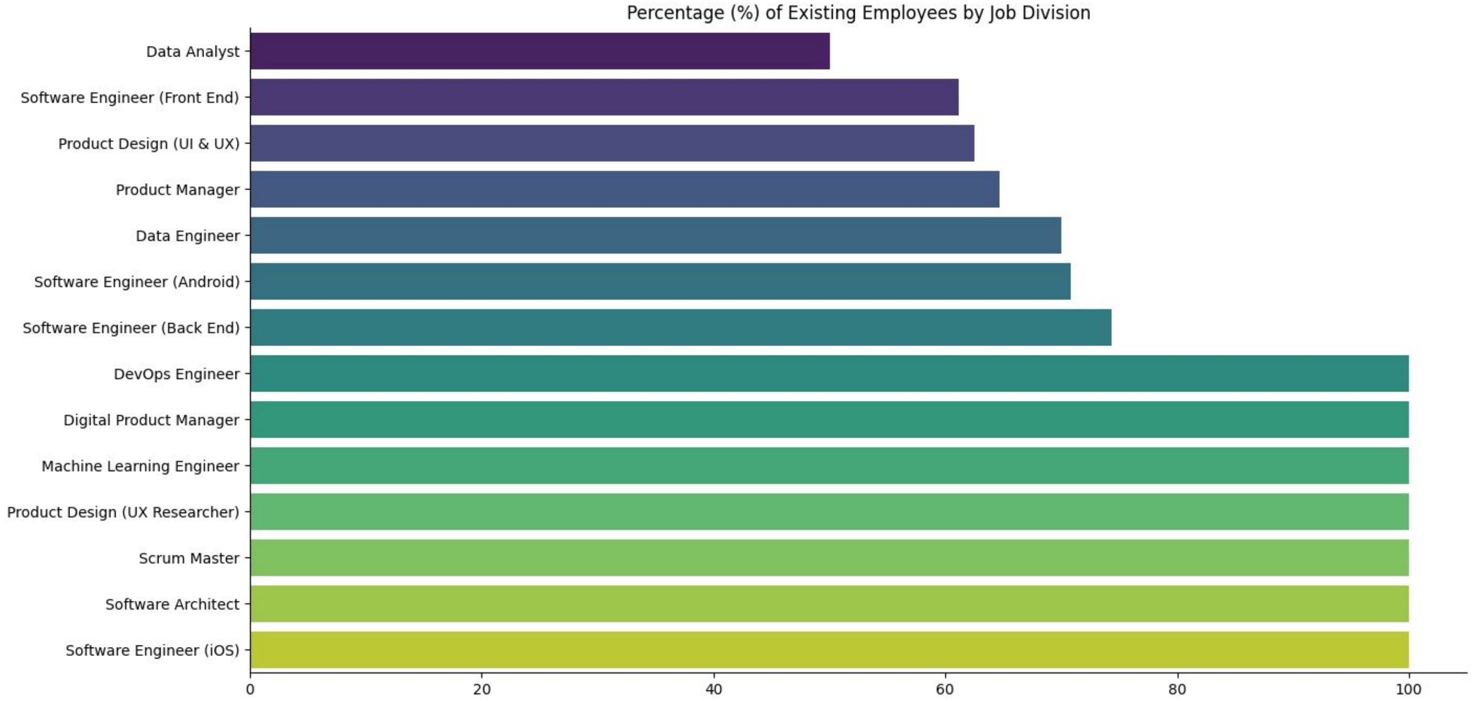
Annual Report on Employee Number Changes



The company has had a total of 287 new hires from 2006 to 2018 and 89 resignations from 2013 to 2020, leaving 198 remaining employees in 2021. The highest number of new hires occurred in 2011, while the highest number of resignations happened in 2018. Additionally, the company has reduced its recruitment of new employees since 2015 and had a hiring freeze since 2019. There was also a high turnover rate in 2017-2018, possibly due to poor financial conditions that require further investigation.

For more details, you can see all file here and code here

Resign Reason Analysis for Employee Attrition Management Strategy

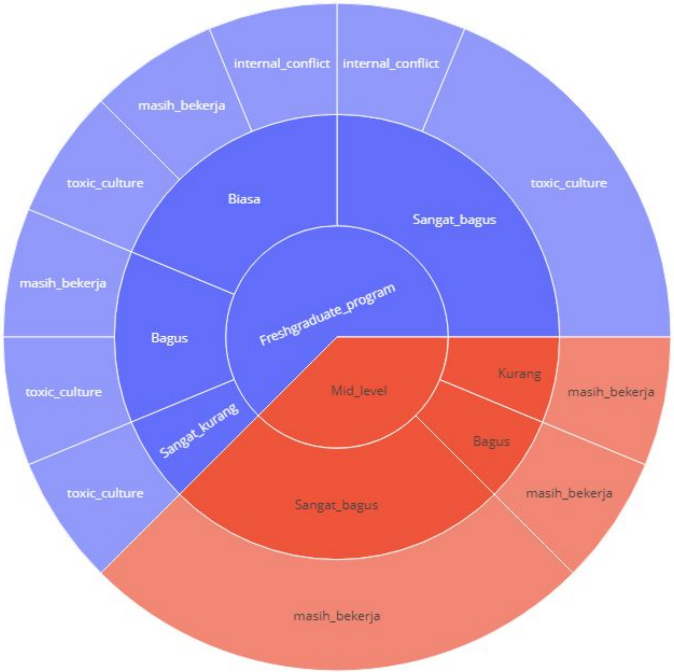


From the viz above, we know that Data Analyst Role is the division with the highest resign rate is the Data Analyst Role with a 50% Existing Employee rate. Overall, the data suggests that the company may need to investigate and address the reasons behind the high resign rates in these divisions to retain their employees and maintain business performance.

For more details, you can see all file here and code here

Resign Reason Analysis for Employee Attrition Management Strategy

Resigned Employees from Data Analyst Division Based on Career Path, Performance, and Reasons for Resignation



The data shows that most of the resigned employees were those with a "Sangat bagus" performance rating and the main reason for resigning was due to a "toxic culture". Additionally, the majority of the resigned employees were fresh graduates with a "Sangat bagus" performance rating and the main reason for resigning was also due to a "toxic culture". There were also a few mid-level employees who resigned, with one having a "Bagus" performance rating and the other two having "Sangat bagus" performance ratings. All mid-level employees resigned while still working.

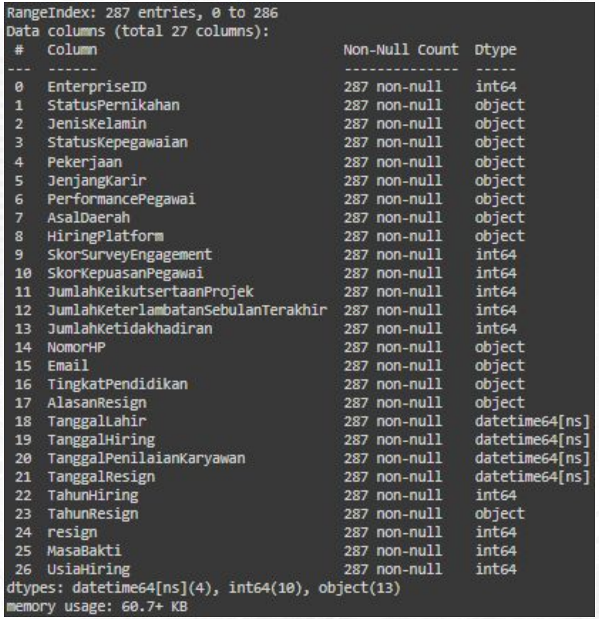
For more details, you can see all file here and code here

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Data Preprocessing

- **Feature Selection**



```
RangeIndex: 287 entries, 0 to 286
Data columns (total 27 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   EnterpriseID                      287 non-null    int64
 1   StatusPernikahan                  287 non-null    object
 2   JenisKelamin                      287 non-null    object
 3   StatusKepegawaian                 287 non-null    object
 4   Pekerjaan                         287 non-null    object
 5   JenjangKarir                      287 non-null    object
 6   PerformancePegawai                287 non-null    object
 7   AsalDaerah                        287 non-null    object
 8   HiringPlatform                    287 non-null    object
 9   SkorSurveyEngagement              287 non-null    int64
 10  SkorKepuasanPegawai               287 non-null    int64
 11  JumlahKeikutsertaanProjek         287 non-null    int64
 12  JumlahKeterlambatanSebulanTerakhir 287 non-null   int64
 13  JumlahKetidakhadiran              287 non-null    int64
 14  NomorHP                           287 non-null    object
 15  Email                             287 non-null    object
 16  TingkatPendidikan                 287 non-null    object
 17  AlasanResign                      287 non-null    object
 18  TanggalLahir                      287 non-null    datetime64[ns]
 19  TanggalHiring                     287 non-null    datetime64[ns]
 20  TanggalPenilaianKaryawan          287 non-null    datetime64[ns]
 21  TanggalResign                     287 non-null    datetime64[ns]
 22  TahunHiring                       287 non-null    int64
 23  TahunResign                       287 non-null    object
 24  resign                            287 non-null    int64
 25  MasaBakti                         287 non-null    int64
 26  UsiaHiring                        287 non-null    int64
dtypes: datetime64[ns](4), int64(10), object(13)
memory usage: 60.7+ KB
```

- **Handling Outlier**

  'JumlahKetidakhadiran', 'resign', 'MasaBakti', 'UsiaHiring', 'SkorKepuasanPegawai'

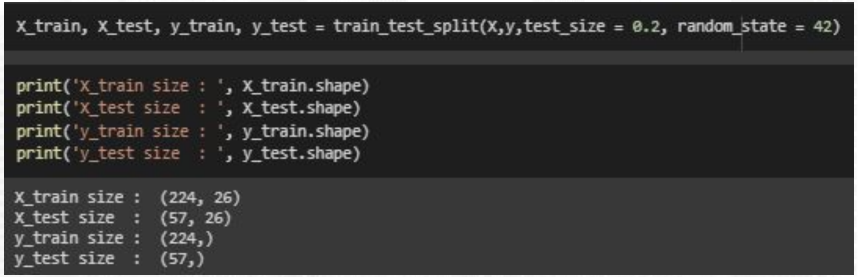
- **Feature Encoding**

  Label Encode, One-Hot Encode, Frequency Encode

For more details, you can see all file here and code here

Split Data & Handling Imbalance

- **Splitting Data**

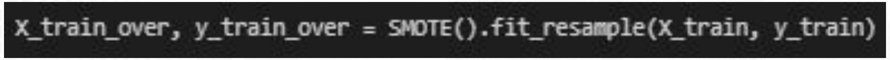


```
x_train, x_test, y_train, y_test = train_test_split(X,y,test_size = 0.2, random_state = 42)
```

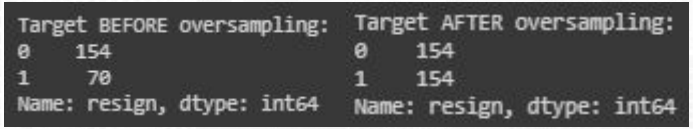
```
print('X_train size : ', X_train.shape)
print('X_test size  : ', X_test.shape)
print('y_train size : ', y_train.shape)
print('y_test size  : ', y_test.shape)
```

```
X_train size :  (224, 26)
X_test size  :  (57, 26)
y_train size :  (224,)
y_test size  :  (57,)
```

- **Handling Imbalance**



```
X_train_over, y_train_over = SMOTE().fit_resample(X_train, y_train)
```



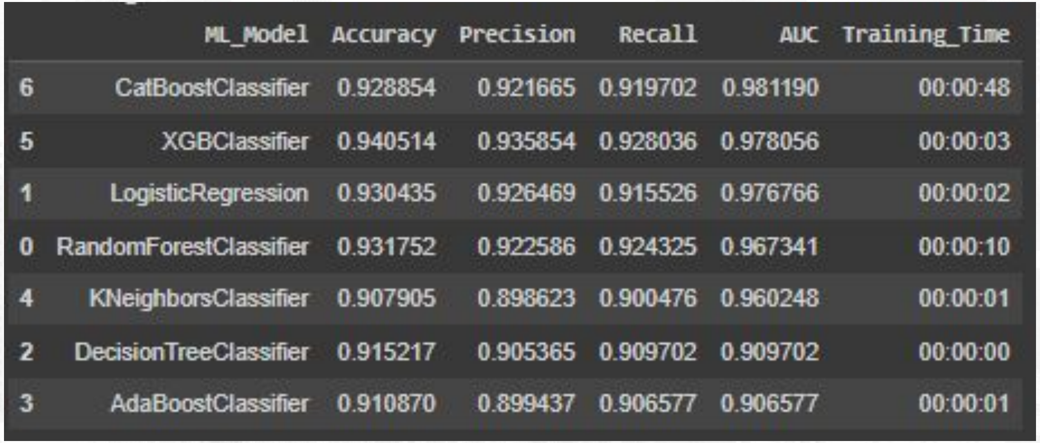
```
Target BEFORE oversampling:    Target AFTER oversampling:
0    154                       0    154
1     70                       1    154
Name: resign, dtype: int64     Name: resign, dtype: int64
```

The dataset was split into a training set and a testing set with an 80:20 ratio, and to handle the imbalance in the target feature 'resign', the SMOTE sampling method was used as a sampling strategy.

For more details, you can see all file here and code here

Build an Automated Resignation Behavior Prediction using Machine Learning

Modelling



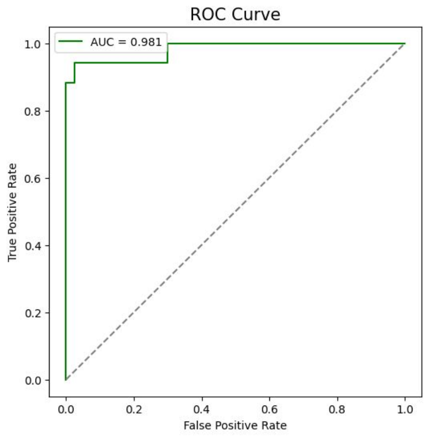
	ML_Model	Accuracy	Precision	Recall	AUC	Training_Time
6	CatBoostClassifier	0.928854	0.921665	0.919702	0.981190	00:00:48
5	XGBClassifier	0.940514	0.935854	0.928036	0.978056	00:00:03
1	LogisticRegression	0.930435	0.926469	0.915526	0.976766	00:00:02
0	RandomForestClassifier	0.931752	0.922586	0.924325	0.967341	00:00:10
4	KNeighborsClassifier	0.907905	0.898623	0.900476	0.960248	00:00:01
2	DecisionTreeClassifier	0.915217	0.905365	0.909702	0.909702	00:00:00
3	AdaBoostClassifier	0.910870	0.899437	0.906577	0.906577	00:00:01

Based on the results, CatBoost Classifier performed the best and XGBoost Classifier came in second in all evaluation metrics. So, we will use CatBoost Classifier for now.

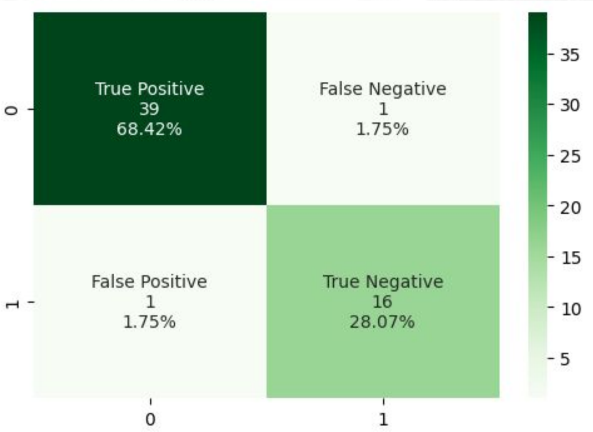
For more details, you can see all file here and code here

Evaluation

ROC Curve



Confusion Matrix



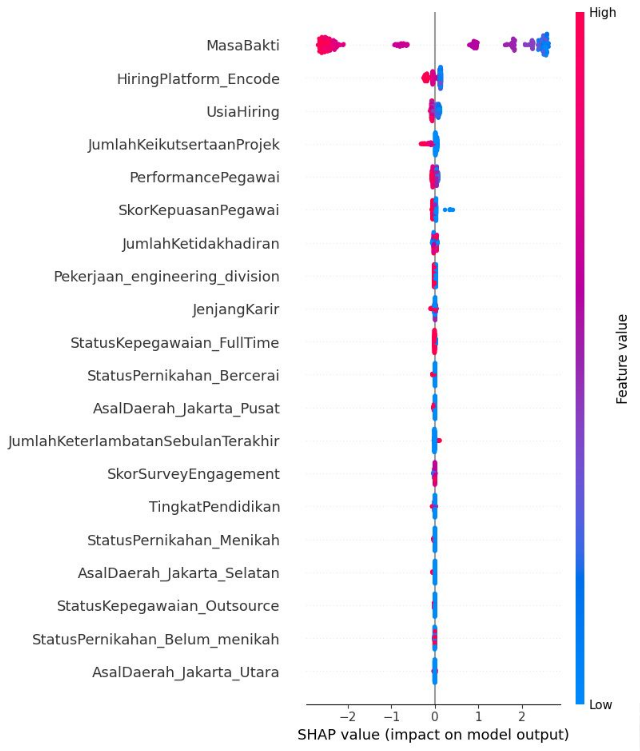
	0	1
0	True Positive 39 68.42%	False Negative 1 1.75%
1	False Positive 1 1.75%	True Negative 16 28.07%

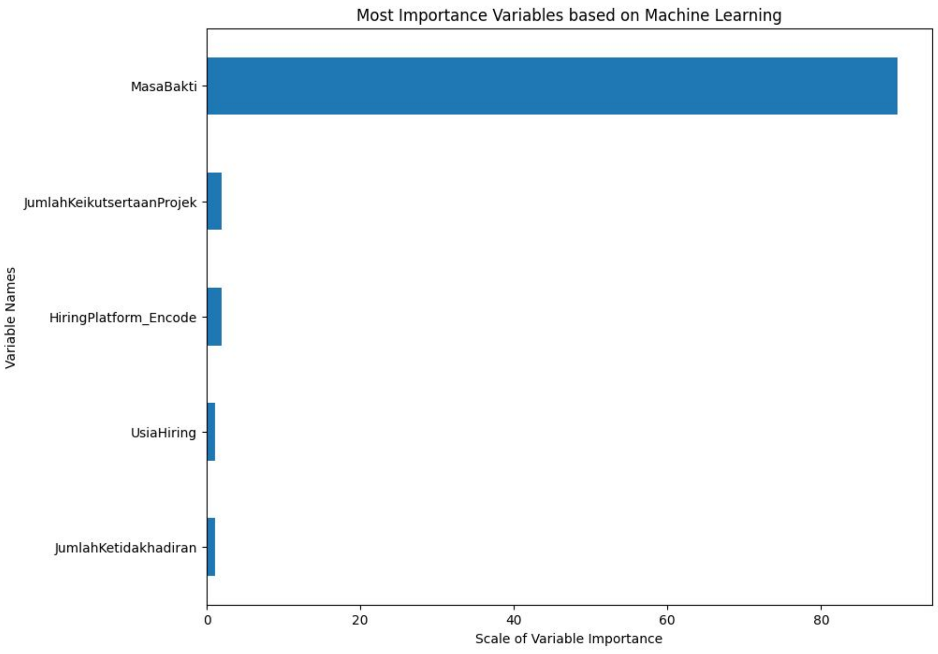
The model which utilized CatBoost Classifier and handled the imbalanced data along with hyperparameter tuning, was able to predict the employees who would stay in the company with very low error.

For more details, you can see all file here and code here

Presenting Machine Learning Products to the Business Users

Feature Importance





For more details, you can see all file here and code here

Presenting Machine Learning Products to the Business Users

Based on the feature importance analysis, it was found that the most influential feature is the length of service, or MasaBakti in Indonesian. Specifically, employees who have worked for less than 6 years are more likely to resign. This trend is consistent with the fact that many of the resigned employees were fresh graduates who had little to no experience and had to adapt to the company's social and professional working environment.

Fresh graduates tend to have idealistic and critical thinking about their job and the work environment, which may influence their performance and decision-making. The partial dependencies plot indicates that the probability of resignation increases from very poor to average levels of employee performance.

The reasons for resignation varied, but the most common ones were related to the company's social culture, such as a toxic working environment, internal conflicts, unhappiness, and lack of appreciation. The most significant reason was related to technical regulations, specifically the lack of flexibility for remote work.

To prevent further performance degradation and financial losses due to employee resignation, a deeper analysis of these reasons should be conducted at every level of the company. New regulations and practices should be developed to enhance the social culture and work habits within the company.

For more details, you can see all file here and code here

Thank You!

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com