Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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

Created by:
Danu Satria Wiratama
danusatria06@gmail.com
https://www.linkedin.com/in/danusatria/

"I am a computer science graduate from Satya Wacana Christian University with a concentration in data science. I completed an online bootcamp at Rakamin Academy with a topic of data science, which further enhanced my skills in data analysis and machine learning. I have mastered several skills such as SQL, python, data visualization, and additional proficiency in machine learning for further modeling and analysis.. "

Overview

“Sumber daya manusia (SDM) adalah aset utama yang perlu dikelola dengan baik oleh perusahaan agar tujuan bisnis dapat tercapai dengan efektif dan efisien. Pada kesempatan kali ini, kita akan menghadapi sebuah permasalahan tentang sumber daya manusia yang ada di perusahaan. Fokus kita adalah untuk mengetahui bagaimana cara menjaga karyawan agar tetap bertahan di perusahaan yang ada saat ini yang dapat mengakibatkan bengkaknya biaya untuk rekrutmen karyawan serta pelatihan untuk mereka yang baru masuk. Dengan mengetahui faktor utama yang menyebabkan karyawan tidak merasa, perusahaan dapat segera menanggulanginya dengan membuat program-program yang relevan dengan permasalahan karyawan. “

Data Preprocessing

- Memahami dan mengeksplorasi data dengan fungsi info dan describe.
- Membagi kolom dengan bentuk numerikal dan kategorikal.
- Melakukan pengecekan data null dan data duplikat, hal ini perlu diperhatikan karena jika terdapat data null akan mempengaruhi hasil analisis dan model prediksi.
- Menghapus kolom ikutprogramLOP.
- Mengisi missing value ‘Lainnya’ pada kolom AlasanResign.
- Mengisi missing value dengan nilai median pada kolom JumlahKetidakhadiran.
- Mengisi missing value dengan nilai median pada kolom SkorKepuasanPegawai.
- Mengisi missing value dengan nilai median pada kolom JumlahKeikutsertaanProyek.
- Mengisi missing value dengan nilai median pada kolom JumlahKeterlambatan.
- Melakukan replace nilai ‘yes’ menjadi ‘1’ pada kolom PernahBekerja.
- Melakukan replace nilai ‘Product Design (UI & UX)’ menjadi nilai mode pada kolom AlasanResign

Data Preprocessing

- Membuat kolom LamaBekerja berdasarkan TanggalResign - TanggalHiring.
- Membuat kolom Usia berdasarkan TanggalLahir
- Membuat kolom Status berdasarkan TanggalResign

Untuk selengkapnya, dapat melihat jupyter notebook disini

Annual Report on Employee Number Changes

Rakamin Academy

Tabel Total Karyawan Setiap Tahun:

	Tahun	JumlahKaryawanMasuk	JumlahKaryawanKeluar	TotalKaryawan
0	2006	1	0	1
1	2007	2	0	3
2	2008	2	0	5
3	2009	7	0	12
4	2010	8	0	20
5	2011	76	0	96
6	2012	41	0	137
7	2013	43	5	175
8	2014	56	12	219
9	2015	31	8	242
10	2016	14	8	248
11	2017	5	19	234
12	2018	1	26	209
13	2019	0	5	204
14	2020	0	6	198

Tabel jumlah karyawan yang masuk / keluar setiap tahunnya beserta total karyawan yang tersedia.

Annual Report on Employee Number Changes

Berdasarkan waterfall chart tersebut terlihat bahwa penurunan pegawai terus terjadi dimulai dari tahun 2016 hingga 2020. Hal ini menunjukkan bahwa kondisi perusahaan sedang tidak sehat dan perlu adanya tindakan lebih lanjut.

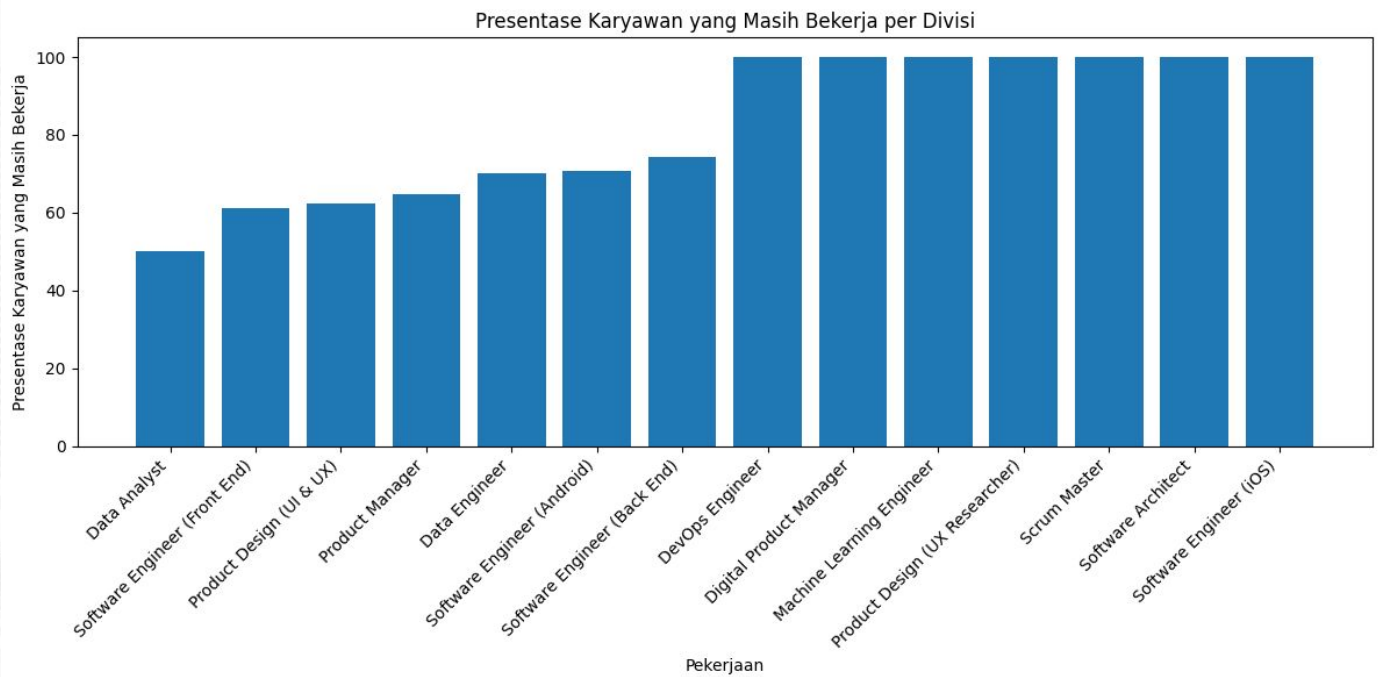
Resign Reason Analysis for Employee Attrition Management Strategy

Rakamin Academy

	Pekerjaan	JumlahKaryawanBekerja	JumlahKaryawanResign
0	Data Analyst	6	6.0
1	Data Engineer	4	2.0
2	Digital Product Manager	2	0.0
3	Machine Learning Engineer	1	0.0
4	Product Design (UI & UX)	11	7.0
5	Product Design (UX Researcher)	1	0.0
6	Product Manager	9	3.0
7	Scrum Master	3	0.0
8	Software Architect	1	0.0
9	Software Engineer (Android)	13	5.0
10	Software Engineer (Back End)	60	17.0
11	Software Engineer (Front End)	26	18.0

Pada tabel tersebut saya mengelompokkan jumlah karyawan yang masih bekerja maupun karyawan yang sudah resign berdasarkan pekerjaannya.

Resign Reason Analysis for Employee Attrition Management Strategy

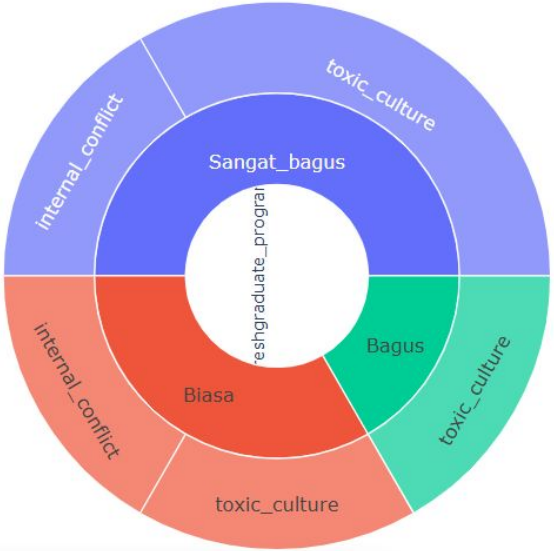


Selanjutnya saya membuat chart untuk menunjukan persentase pekerjaan yang mengalami pengurangan (resign). Dapat dilihat bahwa Data Analyst memiliki tingakat penurunan pegawai yang paling besar.

Resign Reason Analysis for Employee Attrition Management Strategy

Rakamin Academy

Data Analyst Resignations Breakdown



Berdasarkan grafik tersebut kita dapat melihat bahwa dalam divisi data analyst yang melakukan resign, semuanya berasal dari jenjang karir fresh graduates program. Dengan performa antara biasa, bagus, dan sangat bagus. Alasan mereka melakukan resign terdiri dari dua alasan yaitu toxic culture dengan indeks paling tinggi (4 nilai) dan internal konflik (2 nilai).

Hal ini menunjukkan bahwa terdapat ketidaknyamanan dalam budaya kerja (toxic culture) dimana manajemen harus melakukan tindakan lebih lanjut untuk menanggulangi masalah ini, seperti identifikasi sumber masalah dengan survei atau feedback karyawan, mengevaluasi kepemimpinan, atau meningkatkan kesejahteraan karyawan.

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Melakukan check kembali pada data duplikat dan missing value

```
df_clean.isnull().sum()
```

	0
Username	0
EnterpriseID	0
StatusPernikahan	0
JenisKelamin	0
StatusKepegawaian	0
Pekerjaan	0
JenjangKarir	0
PerformancePegawai	0
AsalDaerah	0
HiringPlatform	0
SkorSurveyEngagement	0
SkorKepuasanPegawai	0
JumlahKeikutsertaanProjek	0
JumlahKeterlambatanSebulanTerakhir	0
JumlahKetidakhadiran	0
NomorHP	0
Email	0
TingkatPendidikan	0
PernahBekerja	0
AlasanResign	0
TanggalLahir	0
TanggalHiring	0
TanggalPenilaianKaryawan	0
TanggalResign	198
TahunHiring	0
TahunResign	198
LamaBekerja	0
UsiaHired	0
Status	0

dtype: int64

Masih terdapat null value pada kolom ‘TanggalResign’ dan ‘TahunResign’ namun nantinya kolom tersebut juga akan dihapus sehingga tidak dilakukan pengisian value

```
[ ] df_clean.duplicated().sum()
```

```
0
```

Sudah tidak ditemukan data duplikat

Melakukan feature engineering

- Membuat kolom ‘Status’ berdasarkan kolom ‘TanggalResign’ yang menunjukan bahwa karyawan tersebut telah resign yang bernilai (1), dan jika kolom ‘TanggalResign’ bernilai null maka menujukan bahwa karyawan masih bekerja yang bernilai (0).

```
#Membuat kolom baru bernama 'Status' dimana 1 = resign dan 0 = masih bekerja.
df_clean['Status'] = df_clean.apply(lambda x: 0 if x['TanggalResign']=='-' or pd.isnull(x['TanggalResign']) else 1, axis=1)
df_clean['Status'].value_counts()
```

- Membuat kolom “LamaBekerja” dan “Usia”

```
# Membuat kolom 'LamaBekerja'
df_clean['LamaBekerja'] = (df_clean['TanggalResign'] - df_clean['TanggalHiring']).dt.days
df_clean['LamaBekerja'] = df_clean['LamaBekerja'] // 365  # Convert hari ke tanggal
df_clean['LamaBekerja'] = df_clean['LamaBekerja'].fillna(0).astype(int)

# Membuat kolom 'Usia'
if 'TanggalLahir' in df_clean.columns:
    df_clean['TanggalLahir'] = pd.to_datetime(df_clean['TanggalLahir'], errors='coerce')
    df_clean['UsiaHired'] = (df_clean['TanggalHiring'] - df_clean['TanggalLahir']).dt.days // 365
else:
    print("Error: 'TanggalLahir' tidak ditemukan dalam DataFrame.")
```

Melakukan feature transformation

- Melakukan one hot encoding untuk kolom 'StatusKepegawaian', 'Pekerjaan', 'AsalDaerah', 'HiringPlatform', 'StatusPernikahan', 'AlasanResign'. Dan label encoding pada kolom ‘PerformancePegawai’, ‘TingkatPendidikan’, ‘JenjangKarir’

```
onehot = ['StatusKepegawaian', 'Pekerjaan', 'AsalDaerah', 'HiringPlatform', 'StatusPernikahan', 'AlasanResign']

# Label encoding
df_encoded['PerformancePegawai'] = df_encoded['PerformancePegawai'].map({'Sangat_kurang':  1, 'Kurang': 2, 'Biasa': 3, 'Bagus': 4, 'Sangat_bagus': 5})
df_encoded['TingkatPendidikan'] = df_encoded['TingkatPendidikan'].map({'Sarjana': 1, 'Magister': 2, 'Doktor': 3})
df_encoded['JenjangKarir'] = df_encoded['JenjangKarir'].map({'Freshgraduate_program': 1, 'Mid_level': 2, 'Senior_level': 3})


# One-hot encoding
for cats in onehot:
    onehots = pd.get_dummies(df_encoded[cats], prefix=cats)
    df_encoded = df_encoded.join(onehots)

df_encoded.drop(columns=onehot, axis=1, inplace=True)
```

Melakukan feature transformation

- Melakukan scaling pada kolom 'LamaBekerja', 'UsiaHired', 'SkorSurveyEngagement', 'SkorKepuasanPegawai', 'JumlahKeikutsertaanProjek', ‘JumlahKeterlambatanSebulanTerakhir', 'JumlahKetidakhadiran'

```
from sklearn.preprocessing import MinMaxScaler

# Inisialisasi objek scaler
scaler = MinMaxScaler()

# Melakukan normalisasi Min-Max pada fitur-fitur numerik
numeric_features = ['LamaBekerja', 'UsiaHired', 'SkorSurveyEngagement', 'SkorKepuasanPegawai', 'JumlahKeikutsertaanProjek',
                    'JumlahKeterlambatanSebulanTerakhir', 'JumlahKetidakhadiran']
df_encoded[numeric_features] = scaler.fit_transform(df_encoded[numeric_features])
```

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Melakukan fitur seleksi menggunakan metode chi square dengan pemilihan 15 fitur

```
# Melakukan seleksi fitur dengan fungsi chi-squared dengan k=15
k = 15
selector = SelectKBest(score_func=chi2, k=k)

X = df_encoded.drop(['Status'], axis=1)
y = df_encoded['Status']

X_new = selector.fit_transform(X, y)
selected_feature_indices = selector.get_support(indices=True)
selected_feature_names = X.columns[selected_feature_indices]

print("Selected features:")
print(selected_feature_names)
```

```
Selected features:
Index(['LamaBekerja', 'StatusKepegawaian_Internship', 'Pekerjaan_Data Analyst',
       'StatusPernikahan_-', 'AlasanResign_Lainnya', 'AlasanResign_apresiasi',
       'AlasanResign_ganti_karir', 'AlasanResign_internal_conflict',
       'AlasanResign_jam_kerja', 'AlasanResign_kejelasan_karir',
       'AlasanResign_leadership', 'AlasanResign_masih_bekerja',
       'AlasanResign_tidak_bahagia', 'AlasanResign_tidak_bisa_remote',
       'AlasanResign_toxic_culture'],
      dtype='object')
```

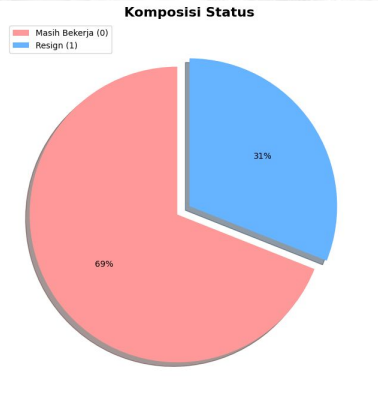
Membagi data menjadi data training dan testing

```python
# Membagi data menjadi data training dan testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

print("Data Split Details:")
print(f"Total Samples: {len(X)}")
print(f"Training Samples: {len(X_train)} ({len(X_train)/len(X)*100:.2f}%)")
print(f"Testing Samples: {len(X_test)} ({len(X_test)/len(X)*100:.2f}%)")
```

```
Data Split Details:
Total Samples: 287
Training Samples: 200 (69.69%)
Testing Samples: 87 (30.31%)
```

Karena target imbalance, dilakukan penyeimbangan data dengan teknik SMOTE




```python
# Menangani data target yang imbalance
from imblearn.over_sampling import SMOTE
smote = SMOTE(random_state=42)
X_train_ro, y_train_ro = smote.fit_resample(X_train, y_train)
```

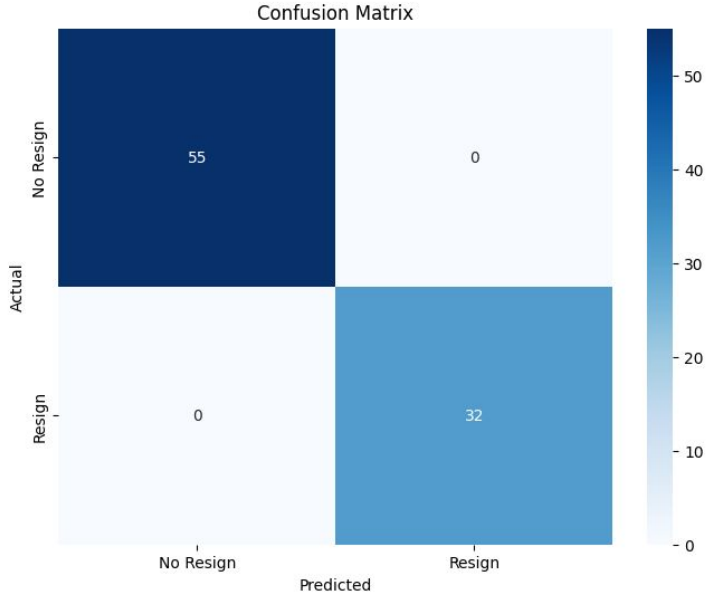
Modelling dengan berbagai metode algoritma

Hasil Evaluasi Model:

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
0	Support Vector Machine	0.977011	1.0	0.93750	0.967742	0.968750
1	Gradient Boosting	1.000000	1.0	1.00000	1.000000	1.000000
2	Decision Tree	0.965517	1.0	0.90625	0.950820	0.953125
3	Random Forest	1.000000	1.0	1.00000	1.000000	1.000000
4	Logistic Regression	0.977011	1.0	0.93750	0.967742	0.968750

Dari berbagai model yang dijalankan, **Gradient Boosting** memiliki nilai Accuracy, Precision, Recall, F1, dan ROC yang paling baik.

Berdasarkan confusion matrix dibawah ini terlihat bahwa model GradientBoosting sangat baik untuk memprediksi data.



Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Hyperparameter

```
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import GradientBoostingClassifier

model = GradientBoostingClassifier(random_state=42)

param_grid = {
    'n_estimators': [50, 100, 150],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [3, 4, 5],
    'min_samples_split': [3, 6, 9],
    'min_samples_leaf': [1, 3, 5],
    'subsample': [0.7, 0.8, 0.9]
}

scoring = 'roc_auc'
grid_search = GridSearchCV(estimator=model, param_grid=param_grid, scoring=scoring, cv=10)
grid_search.fit(X_train, y_train)

# Get the best hyperparameters
best_params = grid_search.best_params_
best_roc_auc = grid_search.best_score_

print("Best Hyperparameters:")
print(best_params)
print(f"Best Mean ROC-AUC: {best_roc_auc:.2f}")
```

```
Best Hyperparameters:
{'learning_rate': 0.05, 'max_depth': 5, 'min_samples_leaf': 3, 'min_samples_split': 3, 'n_estimators': 100, 'subsample': 0.9}
Best Mean ROC-AUC: 0.99
```

Hasil ini menunjukkan bahwa model tersebut menunjukkan daya prediksi yang luar biasa, dengan skor ROC-AUC mencapai sangat tinggi, yaitu 0,99. Pilihan hiperparameter dan skor ROC-AUC yang dihasilkan menggarisbawahi kemampuan model yang kuat untuk membedakan antara kasus positif dan negatif dengan diskriminasi yang sempurna.

Evaluasi dengan cross validation

```python
from sklearn.model_selection import cross_validate
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import make_scorer, roc_auc_score

model = GradientBoostingClassifier()
scoring = {
    'precision': 'precision',
    'recall': 'recall',
    'roc_auc': make_scorer(roc_auc_score)
}

# Membagi data dengan 10 fold
cv_results = cross_validate(model, X_train, y_train, cv=10, scoring=scoring)

# Menghitung mean dan standard deviiasi dari skor precision, recall, dan ROC-AUC
precision_mean = cv_results['test_precision'].mean()
precision_std = cv_results['test_precision'].std()
recall_mean = cv_results['test_recall'].mean()
recall_std = cv_results['test_recall'].std()
roc_auc_mean = cv_results['test_roc_auc'].mean()
roc_auc_std = cv_results['test_roc_auc'].std()

print(f'Mean Precision: {precision_mean:.2f} (±{precision_std:.2f})')
print(f'Mean Recall: {recall_mean:.2f} (±{recall_std:.2f})')
print(f'Mean ROC-AUC: {roc_auc_mean:.2f} (±{roc_auc_std:.2f}')
```

```
Mean Precision: 0.99 (±0.04)
Mean Recall: 0.91 (±0.16)
Mean ROC-AUC: 0.95 (±0.08
```

Hasil ini menunjukkan bahwa model tersebut menunjukkan skor presisi dan perolehan yang kuat, yang menunjukkan kemampuannya untuk mengklasifikasikan kasus positif dengan benar sambil meminimalkan positif palsu dan negatif palsu. Selain itu, skor ROC-AUC yang tinggi mencerminkan kemampuan diskriminasi keseluruhan model yang sangat baik.

Presenting Machine Learning Products to the Business Users

Rakamin Academy

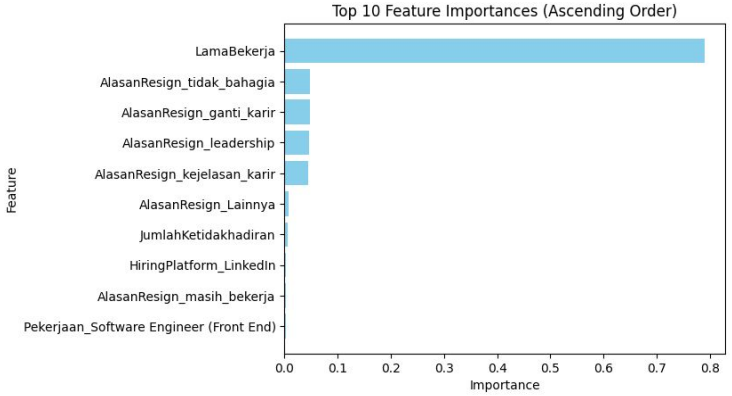
Feature Importance Analysis

- **LamaBekerja**

  Faktor "LamaBekerja" memiliki pengaruh dominan dibandingkan faktor lainnya terhadap variabel target (misalnya churn karyawan, performa, atau outcome terkait lainnya). Dengan nilai kepentingan mendekati **0.8**, ini menunjukkan bahwa durasi seseorang bekerja menjadi faktor kunci dalam analisis ini.

- **AlasanResign_tidak_bahagia**

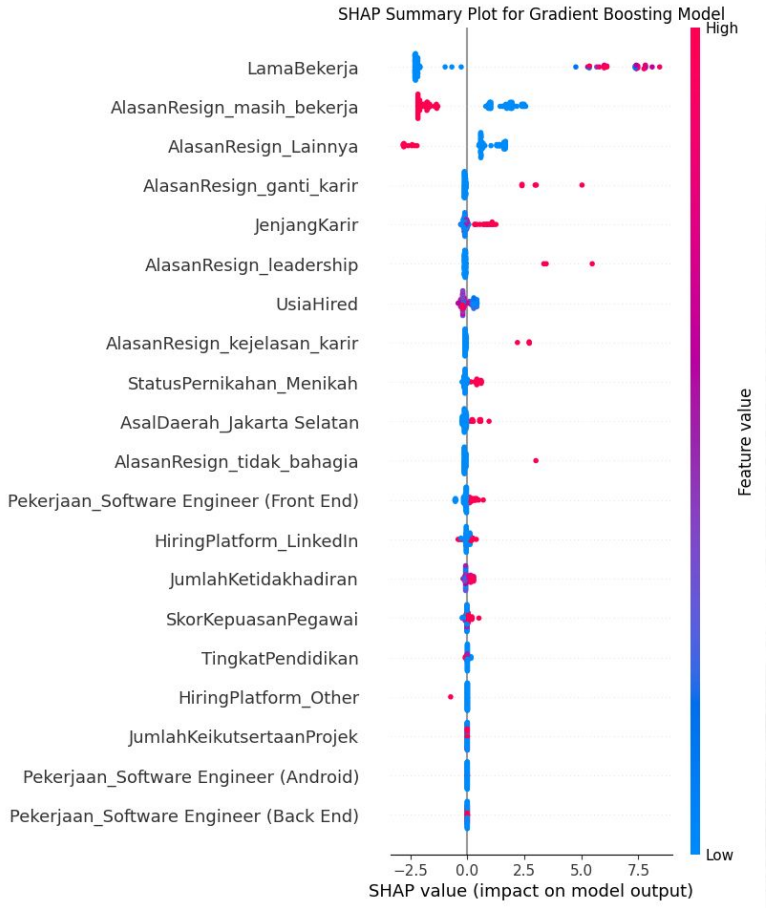
  Berbagai alasan resign menempati faktor penting berikutnya, Hal ini menunjukkan bahwa alasan resign karyawan memiliki kontribusi yang signifikan meskipun tidak sekuat faktor "LamaBekerja".



SHAP Analysis

Berdasarkan analisis SHAP, faktor yang paling mempengaruhi retensi karyawan adalah LamaBekerja. Diikuti dengan faktor lainnya berupa alasan alasan resign seperti ganti karir, leadership, kejelsan karir, dll.

Dengan menangani fitur-fitur utama ini, perusahaan dapat secara proaktif mengurangi pergantian karyawan, meningkatkan kepuasan karyawan, dan pada akhirnya meningkatkan kinerja bisnis mereka.



Business Simulation

Presenting Machine Learning Products to the Business Users

Nama Perusahaan : Eiger Tech

Eiger tech, perusahaan teknologi terdepan, telah berhasil mengembangkan Sistem Peningkatan Retensi Karyawan (SPARK) untuk mengatasi masalah pergantian karyawan yang terus-menerus. Dengan menggunakan teknik pembelajaran mesin yang canggih, SPARK menawarkan solusi komprehensif untuk memprediksi dan meningkatkan retensi karyawan.

Story Telling

Perjalanan Eiger Tech dalam mengembangkan SPARK dimulai dengan misi memberdayakan perusahaan untuk mengurangi pergantian karyawan dan membina tenaga kerja yang lebih stabil dan produktif. Melalui penelitian dan pengembangan selama bertahun-tahun, tim ilmuwan data dan teknisi kami dengan cermat menyusun SPARK untuk menawarkan wawasan dan kekuatan prediktif yang belum pernah ada sebelumnya.

Presenting Machine Learning Products to the Business Users

Wawasan yang kami peroleh dari SPARK sangat berharga bagi bisnis dalam beberapa hal:

1. **LamaBekerja** : Dengan menganalisis pengunduran diri karyawan, SPARK mengidentifikasi bahwa lama bekerja merupakan faktor penting dalam retensi. Hal ini bisa berhubungan dengan berbagai faktor seperti peluang karir yang terbatas, gaji yang stagnan, dan kejenuhan kerja.
2. **AlasanResign_ganti_karir** : SPARK mengungkapkan bahwa alasan resign ganti karir memegang peran penting terhadap retensi. Hal ini menunjukkan bahwa penerapan program atau kebijakan yang bertujuan untuk mempertahankan karyawan jangka panjang sangatlah penting.
3. **JenjangKarir** : Model menunjukan bahwa jenjang karir turut berpengaruh, hal ini menunjukan terdapat perbedaan mendasar di setiap jenjang karir.
4. **AlasanResign_leadership** : Menunjukkan bahwa karyawan merasa tidak puas dengan gaya kepemimpinan atau kualitas pemimpin di perusahaan. Kepemimpinan yang buruk atau tidak efektif dapat berdampak signifikan pada motivasi karyawan.
5. **AlasanResign_kejelasan_karir** : Mengindikasikan bahwa karyawan merasa tidak memiliki arah karir yang jelas atau prospek perkembangan di dalam perusahaan. Hal ini sering terjadi ketika karyawan merasa "terjebak" dalam posisi yang stagnan atau tidak tahu bagaimana langkah mereka selanjutnya dalam organisasi.

Presenting Machine Learning Products to the Business Users

Business Recommendation :

Berdasarkan dari temuan diatas, berikut beberapa rekomendasi untuk mengurangi retensi karyawanL:

1. **Program pengembangan karir** : Memberikan peluang pengembangan karier seperti promosi atau rotasi pekerjaan.
2. **Menciptakan Lingkungan Fleksibel untuk Eksplorasi Skill** : Mengizinkan karyawan untuk bekerja pada proyek lintas fungsi atau inisiatif baru yang sesuai dengan minat mereka.
3. **Menyediakan Jalur Karir yang Transparan** : Buat struktur jenjang karir yang jelas dengan kriteria dan persyaratan yang terukur agar karyawan tahu apa yang harus dicapai untuk naik ke level berikutnya.
4. **Pelatihan Kepemimpinan** : Perusahaan harus memberikan pelatihan khusus kepada pemimpin agar mereka memiliki keterampilan komunikasi, empati, dan manajemen tim yang baik.
5. **Mengadakan Diskusi Karir Secara Rutin** : Mengidentifikasi aspirasi karyawan dan membantu mereka melihat peluang pengembangan dalam perusahaan.

Sistem Peningkatan Retensi Karyawan (SPARK) dari Eiger Tech adalah jawaban untuk mengurangi pergantian karyawan, meningkatkan kepuasan karyawan, dan pada akhirnya meningkatkan kinerja bisnis.

Terimakasih