Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

Created by:
Muchammad Malik
muchammad.malik@gmail.com
https://www.linkedin.com/in/muchammad-malik/

“Currently working as business and system development at GESITS, the leading EV brand in Indonesia. Having a bachelor degree from engineering physics , I possessed balanced skill in engineering and management discipline.

As a data science and business analyst enthusiast, I developed skillset in Business Accumen, SQL, Pyhton, Tableau, and machine learning. I have spent 2 years to learn and maintain these skill by taking several bootcamp and online course.

I am a highly-motivated learner to keep me stay relevant, have good analytical thinking, have creative problem solving skill, and able to work in team. During college life, I developed my soft skill by having experienced in leading a commitee, participating in international competition (Model United Nation and business case competition), and also participating in extra-campus organization, such as AIESEC and StudentsCatalyst. ”

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

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

- Tulislah proses data preprocessing yang kamu lakukan, dan jelaskan secara singkat bagaimana kamu melakukannya, dan alasan mengapa kamu melakukan proses tersebut.

- Source code yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

Data Preprocessing

Handle Missing Value

Username	0
EnterpriseID	0
StatusPernikahan	0
JenisKelamin	0
StatusKepegawaian	0
Pekerjaan	0
JenjangKarir	0
PerformancePegawai	0
AsalDaerah	0
HiringPlatform	0
SkorSurveyEngagement	0
SkorKepuasanPegawai	5
JumlahKeikutsertaanProjek	3
JumlahKeterlambatanSebulanTerakhir	1
JumlahKetidakhadiran	6
NomorHP	0
Email	0
TingkatPendidikan	0
PernahBekerja	0
IkutProgramLOP	258
AlasanResign	66
TanggalLahir	0
TanggalHiring	0
TanggalPenilaianKaryawan	0
TanggalResign	0

1. SkorKepuasanPegawai
2. JumlahKeikutsertaanProjek
3. JumlahKeterlambatanSebulanTerakhir
4. JumlahKetidakhadiran

Missing value diisi dengan nilai rata-rata kolom tersebut (mean) karena data bersifat numerik

5. AlasanResign → Missing value diisi dengan nilai terbanyak (modus) karena data bersifat kategorikal
6. IkutProgramLOP → Kolom tersebut dihapus karena jumlah missing value terlalu banyak

Mengganti value yang tidak Sesuai

```
df['PernahBekerja'].value_counts()
```

```
1      286
yes      1
Name: PernahBekerja, dtype: int64
```

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Data Preprocessing

Mengganti value yang tidak Sesuai

```
df['PernahBekerja'].value_counts()
```

```
1      286
yes      1
Name: PernahBekerja, dtype: int64
```

Nilai ‘yes’ diganti menjadi ‘1’ karena yes dan 1 memiliki sama-sama menunjukkan bahwa seseorang pernah bekerja

```
df['PernahBekerja'].value_counts()
```

```
1    287
Name: PernahBekerja, dtype: int64
```

Membuang data yang tidak diperlukan

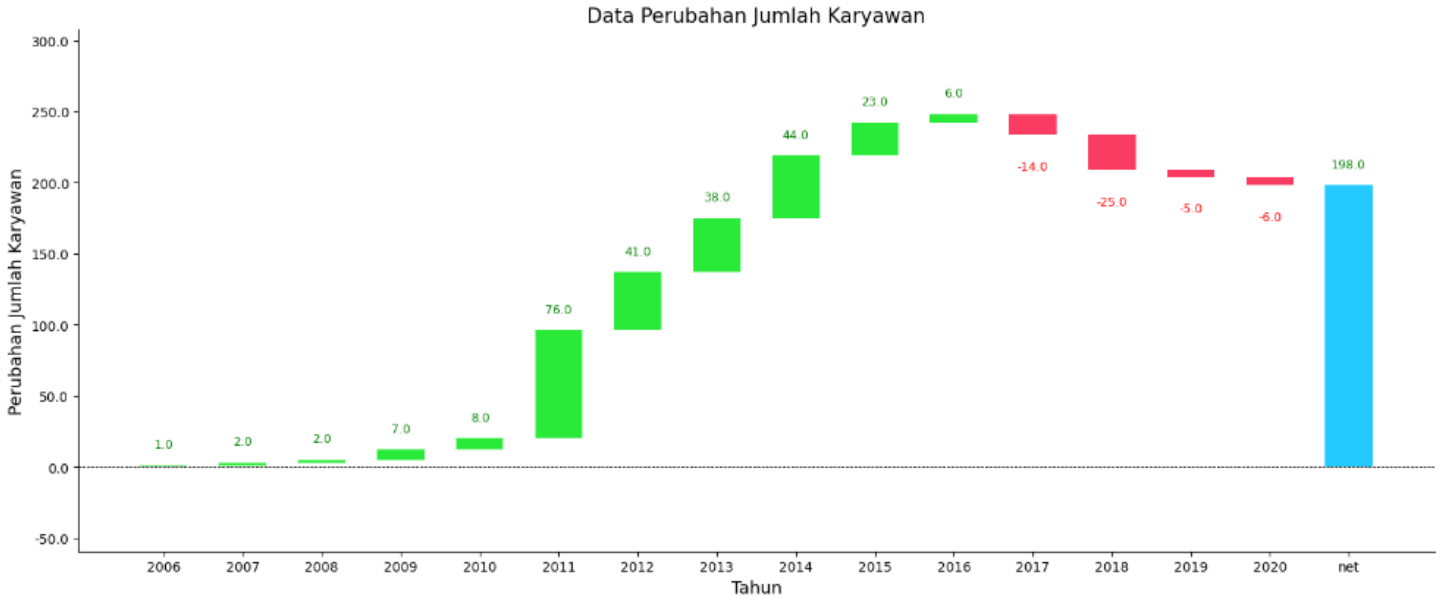
- Data yang tidak diperlukan adalah kolom yang hanya memiliki satu unique value (konstanta)
- Kolom yang hanya memiliki satu unique value adalah PernahBekerja, sehingga kolom tersebut perlu dihapus

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Annual Report on Employee Number Changes

Masukkan grafik visualisasi pada tugas ini, kemudian tuliskan pula hasil analisismu, insight apa saja yang kamu dapatkan.

Annual Report on Employee Number Changes



Plot diatas menunjukkan bahwa kondisi perusahaan sedang menurun akhir-akhir ini

- Hal ini terlihat dari jumlah karyawan yang terus menurun pada periode 2017-2020
- Sementara dari tahun 2006-2016, jumlah karyawan swelalu mengalami kenaikan, dengan kenaikan tertinggi terjadi pada tahun 2011

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

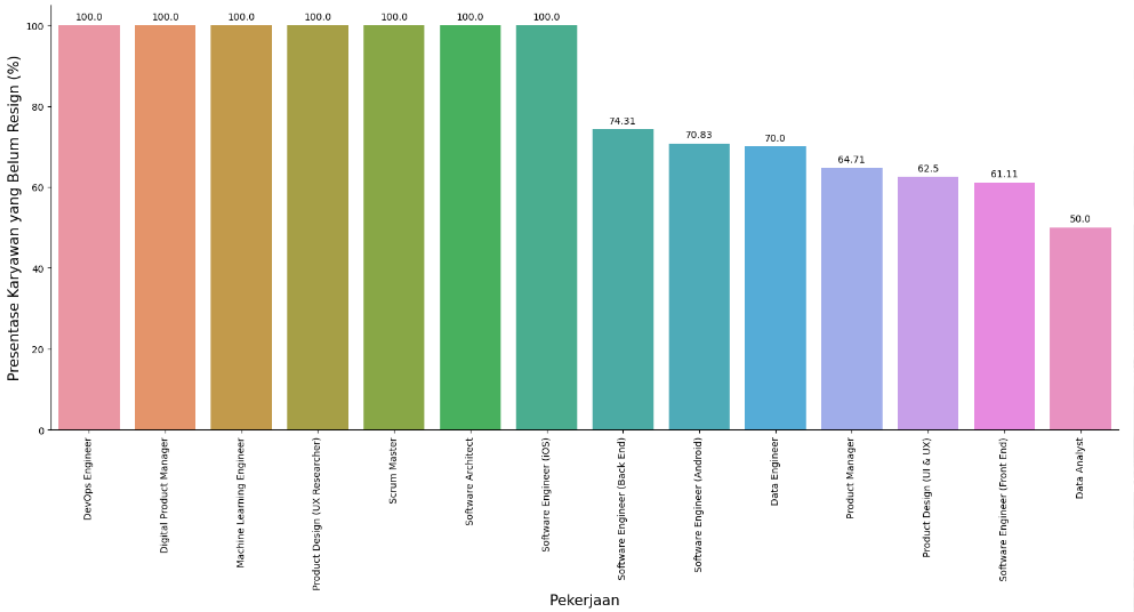
Resign Reason Analysis for Employee Attrition Management Strategy

Rakamin Academy

Masukkan tabel / grafik visualisasi pada tugas ini, kemudian tuliskan pula hasil analisismu, insight apa saja yang kamu dapatkan untuk direkomendasikan kepada perusahaan tentang temuanmu.

Resign Reason Analysis for Employee Attrition Management Strategy

Grafik Presentase Karyawan yang Belum Resign



Pekerjaan dengan karyawan belum pernah resign sama sekali terdapat pada 6 divisi, yaitu DevOps Engineer, Digital Product Engineer, Machine Learning Engineer, UX Researcher, Scrum Master, Software Architect, dan Software Engineer iOS, yaitu sebanyak 0%

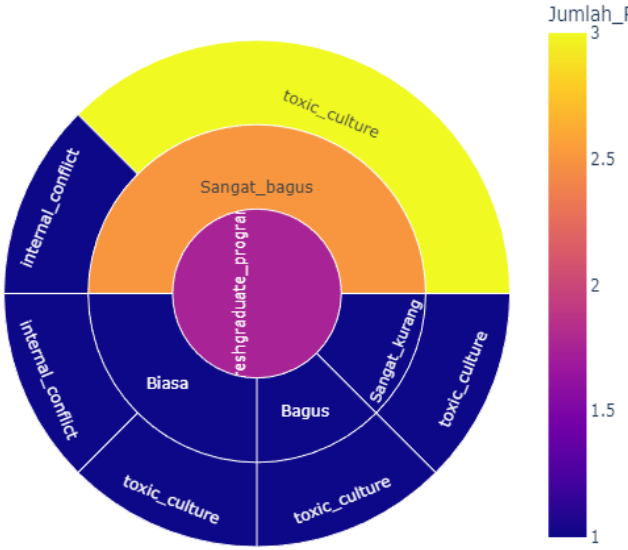
Pekerjaaan dengan tingkat resign paling sedikit adalah Software Engineer, yaitu sebanyak 74,31% karyawan belum resign

Pekerjaaan dengan tingkat resign terbanyak adalah **data analyst**, yaitu sebanyak 50% karyawan sudah resign. Analisa penyebab resign pada pekerjaan ini akan dijelaskan pada slide berikutnya

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Resign Reason Analysis for Employee Attrition Management Strategy

Grafik Jenjang Karir, Performa, dan Alasan Resign pada pekerjaan Data Analyst



- Jenjang karir yang resign pada divisi data analyst hanyalah Fresh Graduate Program
- Sebagian besar karyawan yang resign justru memiliki predikat performa 'Sangat Bagus'. Kemungkinan orang-orang yang memiliki performa bagus diminta mengerjakan tugas tambahan tanpa adanya apresiasi yang semestinya
- Seluruh alasan karyawan resign berhubungan dengan kondisi internal perusahaan, 6 diantaranya menjawab **'toxic culture'** sedangkan 2 orang lagi menjawab **'konflik internal'**

Interpretasi

Kemungkinan terbesar penyebab tingginya tingkat resign pada divisi ini adalah **senior kurang bisa menghargai junior freshgraduate nya**. Hal ini ditandai dengan tidak adanya karyawan senior yang resign, sementara kebanyakan orang yang resign mengatakan internal perusahaan tidak sehat

Tindak Lanjut

Hal yang bias dilakukan manajemen terkait hal ini adalah melakukan assessment berupa Work Load Analysis (WLA) dan cara komunikasi senior dengan junior untuk memastikan beberapa hal berikut:

- Apakah beban kerja tim data analyst masih didalam batas wajar?
- Apakah pembagian tugas antara senior dan junior dalam tim data analyst sudah berimbang?
- Manakah yang perlu diprioritaskan manajemen? Menambah orang atau mengurangi jobdesk?

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Build an Automated Resignation Behavior Prediction using Machine Learning

- Check kembali hasil preprocessing apakah outlier, duplikasi dan missing data telah ter-handle dengan benar
- Melakukan feature transformation dan feature engineering agar data menjadi siap untuk dilakukan modelling
- Melakukan Split data train dan testing lalu dilakukan modelling dengan menggunakan metode machine learning yang berbeda
- Lakukan evaluation dan bandingkan metode machine learning mana yang memiliki hasil terbaik dengan membuat tabel perbandingan

Build an Automated Resignation Behavior Prediction using Machine Learning

Tahapan Preprocessing

Feature Extraction → **Feature Selection** → **Handle Missing dan duplicated value** → **Handle Outlier**

Menambah fitur baru:

- **Resign** : diperoleh dari kolom ‘AlasanResign’
- **Usia** : diperoleh dari kolom ‘TanggalLahir’

Menghapus fitur yang kurang penting untuk proses modelling, seperti Username, EnterpriseID, Email, NomorHP, dsb

```
Jumlah Missing Value:
StatusPernikahan                    0
StatusKepegawaian                   0
Pekerjaan                           0
JenjangKarir                        0
PerformancePegawai                  0
HiringPlatform                      0
SkorSurveyEngagement                0
SkorKepuasanPegawai                 0
JumlahKeikutsertaanProjek           0
JumlahKeterlambatanSebulanTerakhir  0
JumlahKetidakhadiran                0
TingkatPendidikan                   0
TahunHiring                         0
Resign                              0
Usia                                0
dtype: int64

Jumlah Duplicated Value:
0
```

Data yang hilang dan duplikat sudah tidak ada

Tidak semua outlier dihapus karena terdapat beberapa kolom yang nilai outliernya berisi informasi yang bernilai

Outlier dihapus dengan menggunakan IQR

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Build an Automated Resignation Behavior Prediction using Machine Learning

Tahapan Preprocessing

Feature Encoding → **One-Hot Encoding** → **Feature Transformation** → **Handle Imbalance Class**

Kolom yang dilakukan feature encoding:
- **Status Pernikahan**
- **Tingkat Pendidikan**
- **Performance Pegawai**

Kolom yang dilakukan one-hot encoding:
- **Status Kepegawaian**
- **Pekerjaan**
- **Jenjang Karir**
- **Hiring Platform**

Data train dan data test dilakukan normalisasi dengan Min-max Scaler

Dilakukan tiga eksperimen handle imbalance class:
- Undersampling
- Random Oversampling
- Oversampling SMOTE

Split Data train dan data test

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Machine Learning Modelling

Tanpa Oversampling atau Undersampling

	model_name	model	Accuracy (Train Set)	Accuracy (Test Set)	Precision (Train Set)	Precision (Test Set)	Recall (Train Set)	Recall (Test Set)	duration
0	Logistic Regression	LogisticRegression()	0.744395	0.625000	0.769231	0.125000	0.281690	0.066667	0.006001
1	Decision Tree	DecisionTreeClassifier()	1.000000	0.571429	1.000000	0.320000	1.000000	0.533333	0.003001
2	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	1.000000	0.660714	1.000000	0.250000	1.000000	0.133333	0.104938
3	K-Nearest Neighbor	KNeighborsClassifier()	0.753363	0.553571	0.750000	0.083333	0.338028	0.066667	0.001000
4	AdaBoost Classifier	(DecisionTreeClassifier(max_depth=1, random_st...	0.811659	0.589286	0.773585	0.214286	0.577465	0.200000	0.062503
5	GradientBoosting Classifier	([DecisionTreeRegressor(criterion='friedman_ms...	0.968610	0.642857	0.984848	0.222222	0.915493	0.133333	0.062504

Dengan Undersampling

	model_name	model	Accuracy (Train Set)	Accuracy (Test Set)	Precision (Train Set)	Precision (Test Set)	Recall (Train Set)	Recall (Test Set)	duration
0	Logistic Regression	LogisticRegression()	0.704225	0.357143	0.698630	0.161290	0.718310	0.333333	0.007001
1	Decision Tree	DecisionTreeClassifier()	1.000000	0.482143	1.000000	0.181818	1.000000	0.266667	0.001000
2	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	1.000000	0.482143	1.000000	0.250000	1.000000	0.466667	0.080429
3	K-Nearest Neighbor	KNeighborsClassifier()	0.725352	0.517857	0.758065	0.227273	0.661972	0.333333	0.000000
4	AdaBoost Classifier	(DecisionTreeClassifier(max_depth=1, random_st...	0.816901	0.446429	0.835821	0.214286	0.788732	0.400000	0.046860
5	GradientBoosting Classifier	([DecisionTreeRegressor(criterion='friedman_ms...	0.992958	0.446429	1.000000	0.192308	0.985915	0.333333	0.031253

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Machine Learning Modelling

Dengan Random Oversampling

	model_name	model	Accuracy (Train Set)	Accuracy (Test Set)	Precision (Train Set)	Precision (Test Set)	Recall (Train Set)	Recall (Test Set)	duration
0	Logistic Regression	LogisticRegression()	0.723684	0.482143	0.720779	0.250000	0.730263	0.466667	0.009003
1	Decision Tree	DecisionTreeClassifier()	1.000000	0.553571	1.000000	0.272727	1.000000	0.400000	0.003001
2	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	1.000000	0.642857	1.000000	0.272727	1.000000	0.200000	0.108946
3	K-Nearest Neighbor	KNeighborsClassifier()	0.763158	0.517857	0.740964	0.285714	0.809211	0.533333	0.000000
4	AdaBoost Classifier	(DecisionTreeClassifier(max_depth=1, random_st...	0.792763	0.464286	0.773006	0.200000	0.828947	0.333333	0.046876
5	GradientBoosting Classifier	([DecisionTreeRegressor(criterion='friedman_ms...	0.986842	0.535714	0.986842	0.176471	0.986842	0.200000	0.046878

→ **Model yang dipilih**

Dengan Oversampling SMOTE

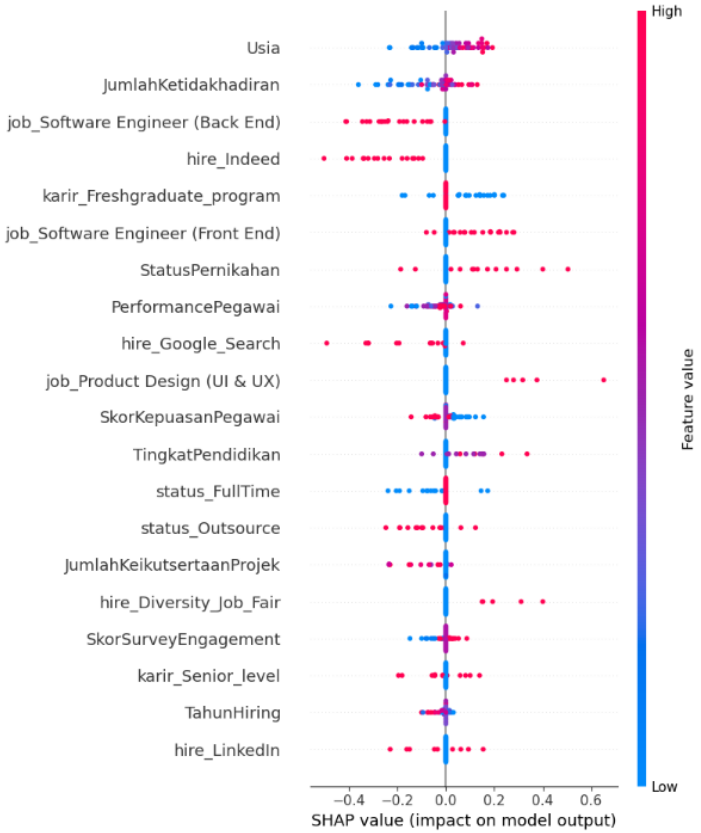
	model_name	model	Accuracy (Train Set)	Accuracy (Test Set)	Precision (Train Set)	Precision (Test Set)	Recall (Train Set)	Recall (Test Set)	duration
0	Logistic Regression	LogisticRegression()	0.667763	0.410714	0.652695	0.178571	0.717105	0.333333	0.008001
1	Decision Tree	DecisionTreeClassifier()	1.000000	0.464286	1.000000	0.200000	1.000000	0.333333	0.002001
2	Random Forest	(DecisionTreeClassifier(max_features='auto', r...	1.000000	0.714286	1.000000	0.454545	1.000000	0.333333	0.100585
3	K-Nearest Neighbor	KNeighborsClassifier()	0.796053	0.392857	0.750000	0.228571	0.888158	0.533333	0.000000
4	AdaBoost Classifier	(DecisionTreeClassifier(max_depth=1, random_st...	0.835526	0.517857	0.835526	0.200000	0.835526	0.266667	0.046877
5	GradientBoosting Classifier	([DecisionTreeRegressor(criterion='friedman_ms...	0.976974	0.607143	0.993197	0.230769	0.960526	0.200000	0.093758

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Presenting Machine Learning Products to the Business Users

Buatlah story telling, makna, dan rekomendasi dari model machine learning yang kamu buat bisa sangat bermanfaat untuk perusahaan dalam perspektif bisnis atau kebutuhan penyelesaian masalah yang ada.

Presenting Machine Learning Products to the Business Users



Interpretasi dari Shap Value

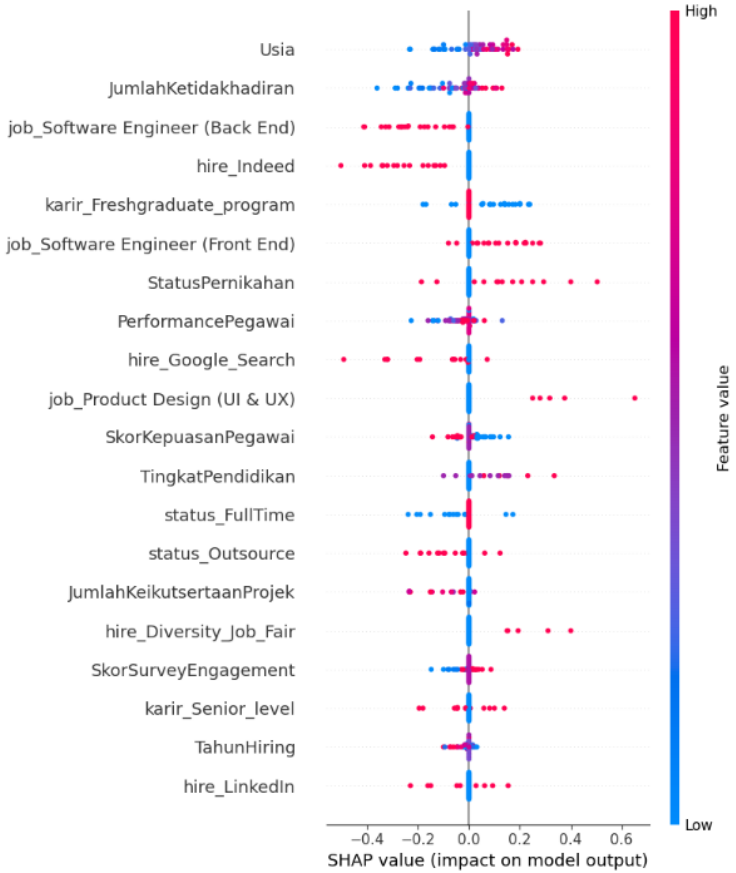
Usia dan Jumlah Ketidakhadiran merupakan dua faktor paling berpengaruh terhadap keputusan resign. Semakin besar nilai dari dua fitur tersebut, semakin besar peluang karyawan resign

Pengaruh Posisi Pekerjaan:
- Software Engineer (Back-End) : Potensi tidak resign sangat tinggi
- Software Engineer (Front-End) : Potensi resign sangat tinggi
- Product Design (UI & UX) : Potensi resign tinggi

Pengaruh Platform Hiring:
- Indeed: Potensi tidak resign sangat tinggi
- Google Search: Potensi tidak resign sangat tinggi
- Diversity job fair: Potensi resign tinggi
- Linkedin: Potensi resign dan tidak resign cukup berimbang

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing

Presenting Machine Learning Products to the Business Users



Rekomendasi Bisnis

Untuk mengurangi jumlah karyawan resign, beberapa langkah yang dapat dilakukan diantaranya:

Terkait Fitur terpenting

- Memprioritaskan rekrutmen karyawan **berusia muda** untuk posisi yang tidak memerlukan pengalaman kerja beberapa tahun
- Melakukan pendekatan personal terhadap karyawan dengan **jumlah ketidakhadiran tinggi** untuk mencari tahu alasan karyawan tersebut merasa tidak nyaman

Terkait Divisi / Pekerjaaan

- Mempelajari pola kerja divisi **back-end engineer** dikarenakan dvisi tersebut memiliki peluang resign yang rendah
- Menganalisa kekurangan pola kerja pada divisi **Front-end engineer dan Product Design (UI & UX)**

Terkait Hiring Platform

- Mengurangi rekrutmen via **Diversity Job Fair** karena platform tersebut memiliki kemungkinan resign yang lebih banyak
- Memperbanyak rekrutmen via **Indeed dan Google** Search karena platform tersebut dapat memperbesar peluang tidak resign

Untuk selengkapnya, dapat melihat pengerjaan disini:
https://colab.research.google.com/drive/1VPHMFk50qQ4OvctLlVM1APNx4d8y_Whm?usp=sharing