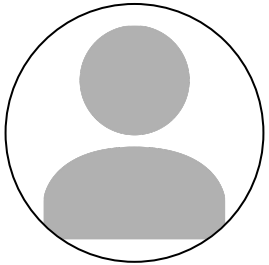
Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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



Created by:
Rika Elisabeth
rikaelisabeth09@gmail.com
https://www.linkedin.com/in/rikaelisabeth/

Data-Driven Geomatics Engineering Graduate Skilled in leveraging data to drive informed decision making, with a strong technical background and diverse internship experiences. Proficient in SQL, Python, and data visualization tools, adept at extracting, transforming, and analyzing complex datasets to uncover meaningful insights. Demonstrated expertise in administrative management, with a keen eye for detail and the ability to manage cross-functional projects effectively. Seeking a challenging role as a Data Analyst where I can apply my analytical acumen and technical expertise to contribute to the organization's success.

Overview

Rakamin Academy

"Human resources (HR) are the main assets that need to be well-managed by a company to achieve business objectives effectively and efficiently. On this occasion, we will face a problem regarding the human resources within the company. Our focus is to find out how to retain employees in the current company to avoid the increased costs of employee recruitment and training for new hires. By identifying the main factors that cause employees to feel dissatisfied, the company can immediately address them by creating programs relevant to employee issues"

Data Preprocessing

Data Preprocessing

Rakamin Academy

Handle Missing and Duplicate Values

Missing Values: Identify and address missing data using imputation techniques (e.g., mean, median, or advanced methods) or by removing rows with missing target values.

Duplicate Values: Find duplicates using df.duplicated() and remove them with df.drop_duplicates().

Drop Unnecessary or Incorrect Values

Identify and Remove: Locate incorrect or irrelevant values and either correct or remove them based on context and analysis needs.

Check Statistical Summary of Categorical and Numerical Features

Categorical Features: Use `df.describe(include=['object'])` to review distributions and frequencies.

Numerical Features: Use `df.describe()` for statistical measures like mean, median, and standard deviation.

Feature Engineering

Create New Features: Generate new features from existing ones (e.g., age and date) and transform features as needed.

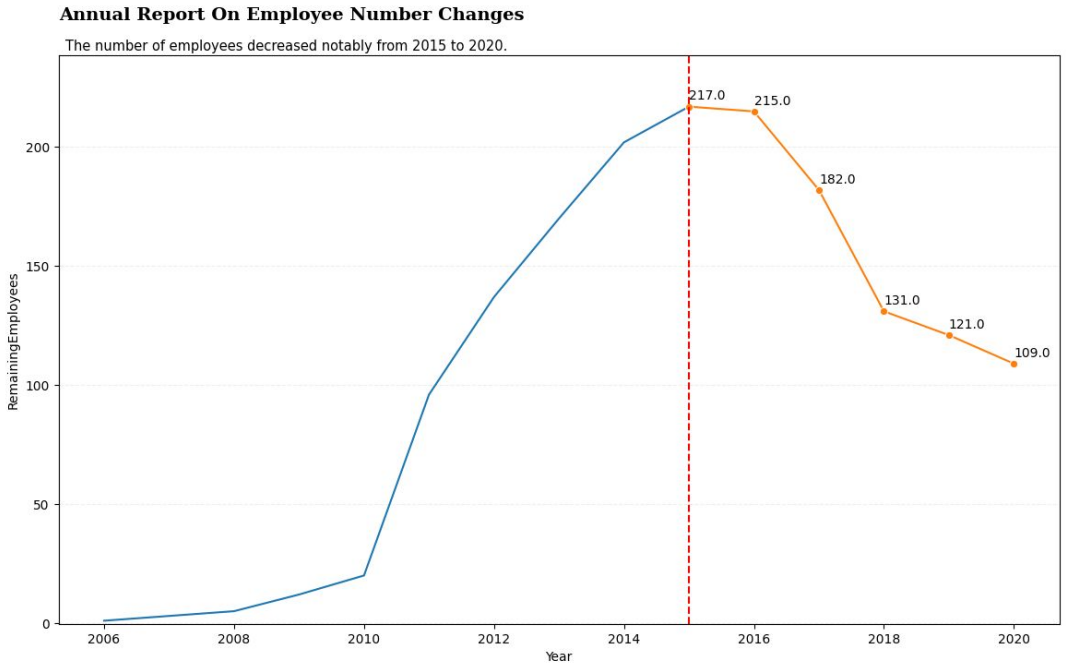
Feature Selection: Choose the most relevant features using techniques for correlation analysis.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Annual Report on Employee Number Changes

Annual Report on Employee Number Changes

The Number of Employees Descreased Notably From 2015 to 2020



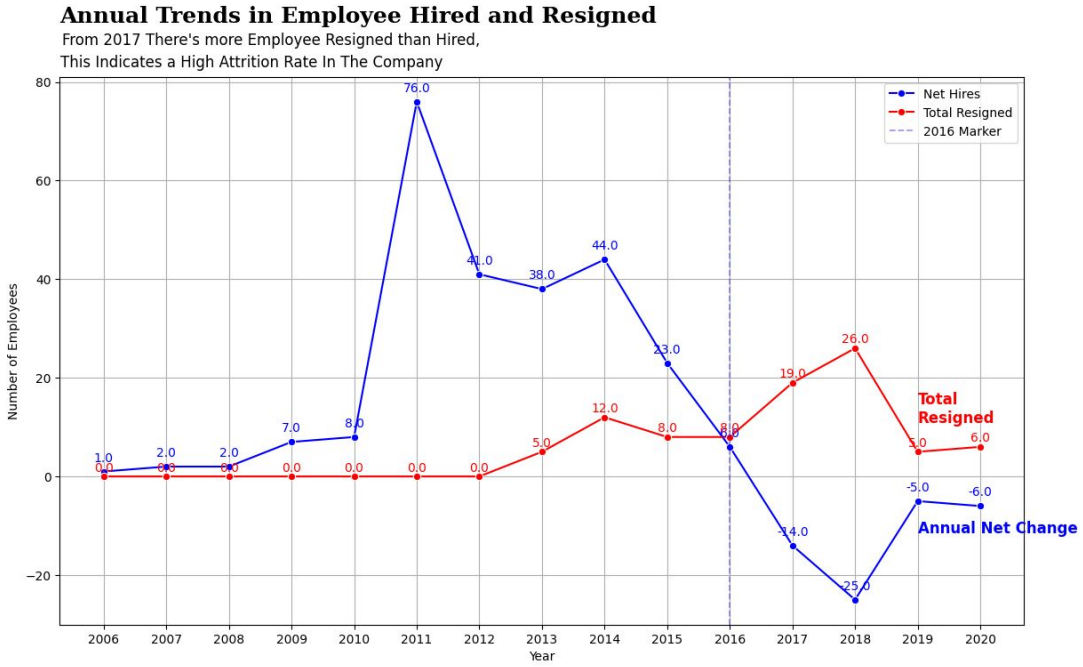
Insight:

The graph shows the total number of employees remaining over the years from 2006 to 2020. The number of employees increased until 2015, when it began to decrease significantly. This decline could indicate a few factors, including economic downturn, company restructuring, or outsourcing. It's important to note that this is just a visual representation and further investigation would be needed to determine the specific causes of the decline.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Annual Report on Employee Number Changes

Annual Trends in Employee Hired and Resigned



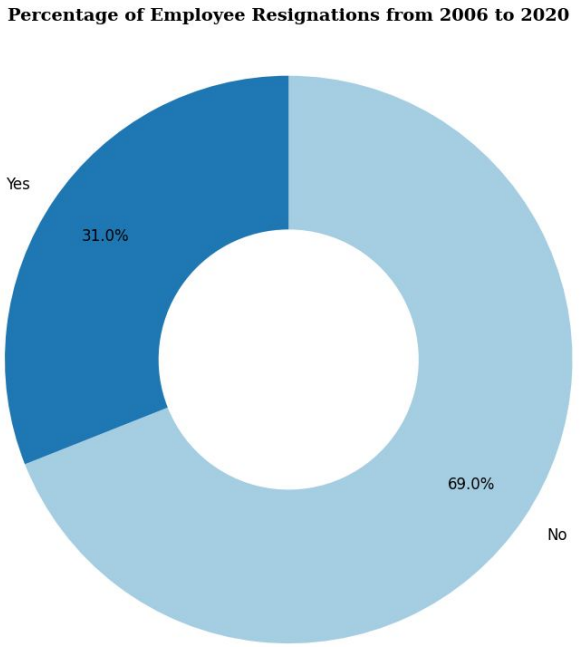
Insight:

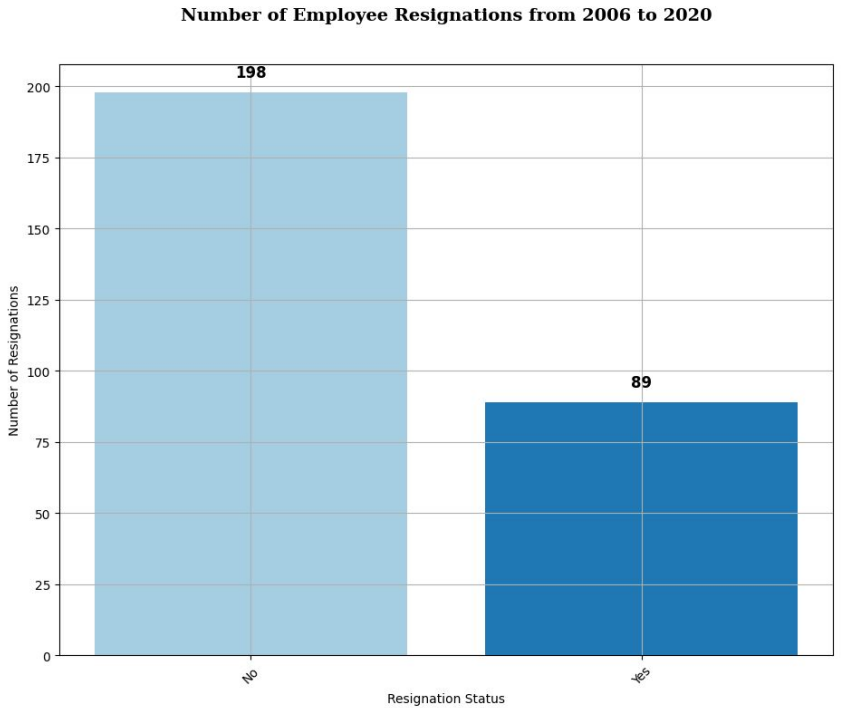
The number of remaining employees over the years, starting from 2006. The graph shows an upward trend from 2006 to 2015, with a sharp decrease starting in 2016. The data suggests that the total number of remaining employees decreased significantly from 2015 to 2020.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Annual Report on Employee Number Changes

Percentage of Employee Resignation From 2006 to 2020





Untuk selengkapnya, dapat melihat jupyter notebook disini

Annual Report on Employee Number Changes

Percentage of Employee Resignation From 2006 to 2020

Insight:

Based on the information provided in the two images, a few key insights can be made:

1. **Percentage of Employee Resignations**: The pie chart shows that 69.0% of employees answered "No" to the question about resignations, while 31.0% answered "Yes". This indicates that a significant majority of employees did not resign during the 2006-2020 period.
2. **Number of Employee Resignations**: The bar chart shows the number of employee resignations over the same time period. The number of "No" resignations is 198, while the number of "Yes" resignations is 89. This reinforces the finding that the majority of employees did not resign.
3. **Trend over Time**: While the data is not broken down by year, the overall trend suggests that employee resignations were relatively low during the 2006-2020 period, with the majority of employees choosing to stay with the organization.

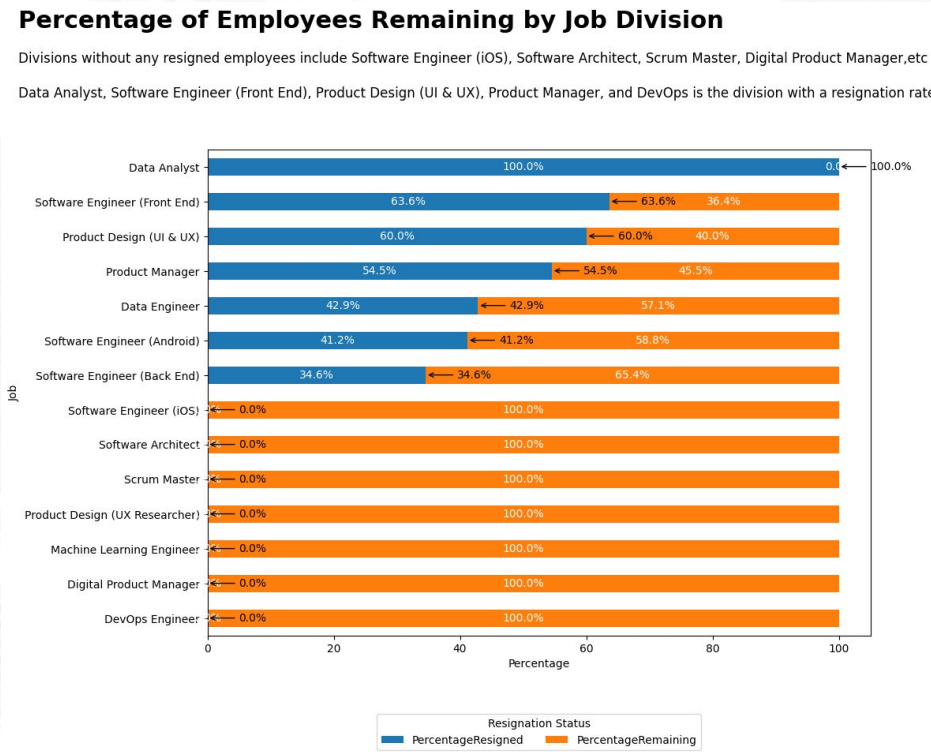
In summary, the data indicates that the vast majority of employees, around 69%, did not resign from their positions between 2006 and 2020, suggesting a relatively stable workforce during that time period.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Resign Reason Analysis for Employee Attrition Management Strategy

Percentage of Employees Remaining by Job Division



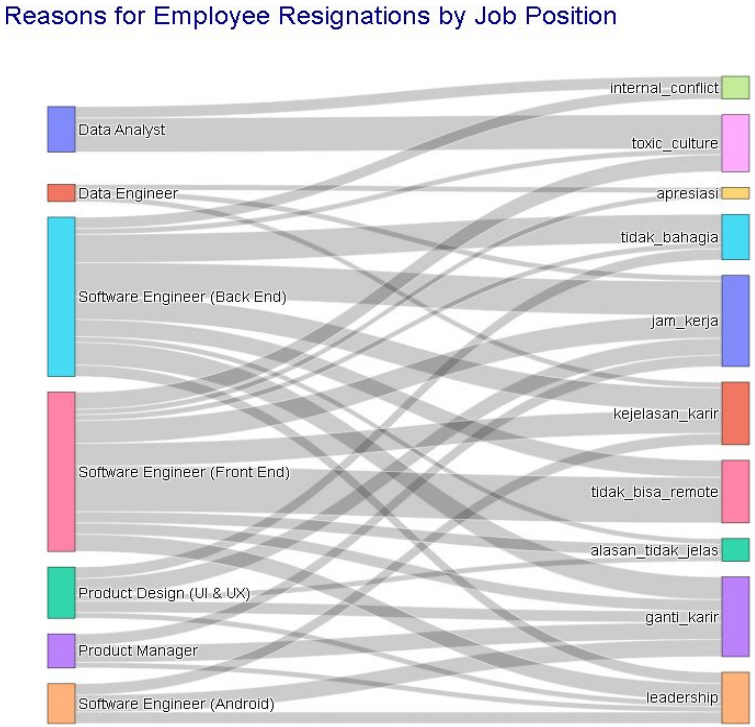
Insight:

1. Divisions with no resigned employees (100% remaining): Software Engineer (iOS), Software Architect, Scrum Master, Digital Product Manager, Product Design (UX Researcher), Machine Learning Engineer, and DevOps Engineer.
2. The division with the highest resignation rate of 50% is Data Analyst, Software Engineer (Front End), Product Design (UI & UX), Product Manager, and DevOps.
3. The divisions with a resignation rate below 50% include Software Engineer (Back End) at 34.6%, Software Engineer (Android) at 41.2%, and Data Engineer at 42.9%.
4. The visualization clearly separates the "`PercentageResigned`" and "`PercentageRemaining`" for each job division, making it easy to identify the divisions with high and low retention rates.
5. This data can help the organization identify job roles that may need more attention or improvements to reduce the high resignation rates and improve employee retention.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Reasons for Employee Resignations by Job Position



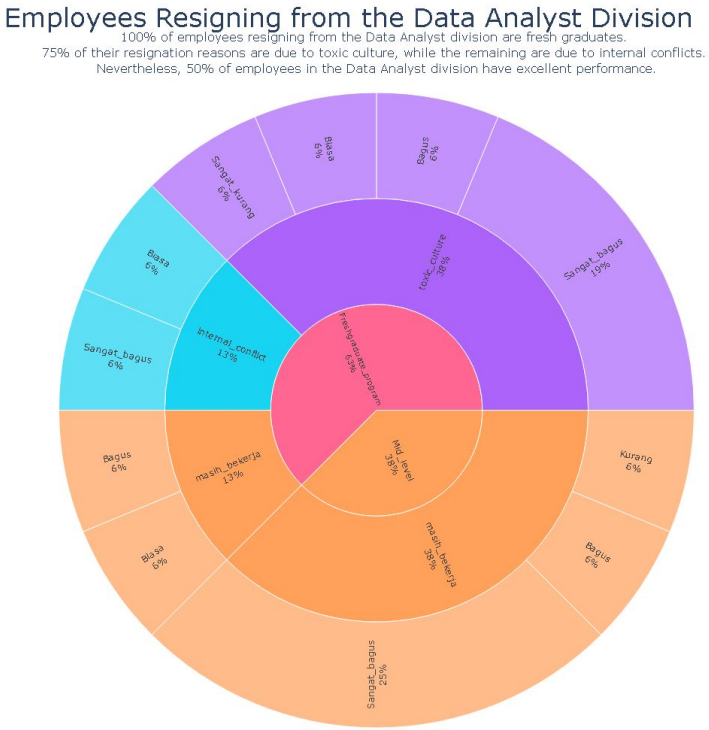
Insight:

1. **Data Analyst Position**: For the Data Analyst position, "toxic culture" accounts for 75% of the resignation reasons, and "internal conflict" accounts for 15%. This indicates that the Data Analyst division is facing a significant challenge in its work culture and internal dynamics, with 90% of resignations attributed to these two factors. Toxic culture refers to a harmful work environment where negative behaviors, such as bullying, lack of support, and poor communication, prevail. Internal conflict denotes disputes and disagreements among team members that disrupt the workflow and create a hostile atmosphere.
2. **Software Engineer (Back End) Position**: For the Software Engineer (Back End) position, "jam kerja" (work hours) accounts for 50% of the resignation reasons, and "tidak bisa remote" (inability to work remotely) accounts for 30%. This suggests that work-life balance and flexibility in work arrangements are major pain points for the Back End Software Engineers. Work hours refer to the long or irregular working hours that can lead to burnout and dissatisfaction. Inability to work remotely indicates that the lack of remote work options limits employees' flexibility and impacts their work-life balance.
3. **Product Design (UI & UX) Position**:For the Product Design (UI & UX) position, "alasan tidak jelas" (unclear reasons) accounts for 60% of the resignation reasons, and "ganti karir" (change in career) accounts for 40%. The high percentage of unclear reasons and career changes indicates a need for better understanding of the specific challenges and aspirations of the Product Design team. Unclear reasons imply that employees leave without providing specific feedback on their reasons for resignation. Career change refers to employees leaving to pursue different career paths, suggesting dissatisfaction with current career development opportunities.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Employees Resigning from the Data Analyst Division



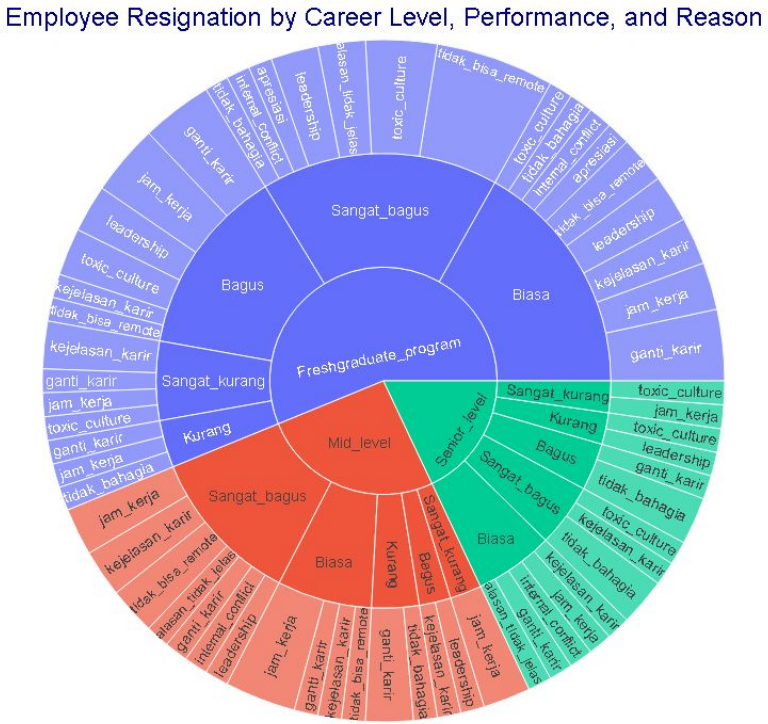
Insight:

1. **100% of the employees resigning from the Data Analyst division are fresh graduates.**
2. **The primary reason for resignations is "toxic_culture", accounting for 75% of the total resignations.**
3. **The second most prevalent reason is "internal_conflict", contributing to 15% of the resignations.**
4. **The remaining 10% of resignations are due to factors like "Biasa" (8%), "Sangat bogus" (6%), and "Kurang" (6%).**
5. **The visualization highlights that all the employees resigning from the Data Analyst division are fresh graduates, indicating a high turnover rate among new hires in this division.**
6. **Despite the high resignation rate, the data notes that 50% of employees in the Data Analyst division have excellent performance, implying that the division is losing some of its top talent due to the organizational and management challenges.**
7. **The visualization provides a comprehensive breakdown of the various reasons for resignations, allowing the organization to identify the root causes and implement targeted interventions to improve the work culture, address internal conflicts, and retain top-performing employees.**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Employee Resignation by Career Level, Performance, and Reason



Insight:

1. **The visualization provides a comprehensive overview of employee resignations across different career levels, performance levels, and reasons for leaving the organization.**
2. **Employees at the "Fresh graduate_program" level are primarily resigning due to "Sangat_kurang" (highly lacking) and "toxic_culture" factors, indicating challenges in the onboarding and retention of new graduates.**
3. **Employees at the "Mid_level" are resigning due to a mix of reasons, including "Sangat_bagus" (highly good), "Sangat_kurang" (highly lacking), and "Biasa" (average), suggesting potential issues with career progression, performance management, and work environment.**
4. **At the "Bagus" (good) performance level, employees are resigning due to "Sangat_kurang" (highly lacking) and "Biasa" (average) factors, implying a disconnect between their performance and the organization's support or recognition.**
5. **The "Biasa" (normal) performance level has a wide range of resignation reasons, including "toxic_culture", "leadership", and "tidak_bisa_remote" (unable to work remotely), highlighting the complexity of issues faced by this group of employees.**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Rakamin Academy

Recommendations:

1. Fresh Graduate Program:
   - Onboarding and Training: Redesign the onboarding program to ensure new graduates receive comprehensive training and support. Introduce a buddy system where new hires are paired with experienced employees for guidance.
   - Address Toxic Culture: Conduct workshops and training sessions focused on fostering a positive work environment and addressing toxic behaviors. Implement regular check-ins with new graduates to address any issues early on.
   - Support Systems: Enhance support systems for new graduates by providing access to resources, career counseling, and mentorship programs.
2. Mid-Level Employees:
   - Career Development: Create clear career progression pathways and provide opportunities for skill development and advancement. Offer training programs, certifications, and rotational assignments to keep employees engaged and growing.
   - Performance Management: Improve the performance management process by providing regular, constructive feedback and recognizing achievements. Address performance issues promptly and provide support for improvement.
   - Work Environment: Foster a collaborative and inclusive work environment. Conduct team-building activities and encourage open communication to address any work environment issues.
3. Good Performance Level Employees:
   - Recognition and Rewards: Develop a robust recognition program that acknowledges and rewards high-performing employees. This can include bonuses, promotions, and public recognition.
   - Challenging Assignments: Provide opportunities for high performers to take on challenging and meaningful projects that align with their career goals.
   - Support and Resources: Ensure high-performing employees have access to the resources and support they need to continue excelling in their roles.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Rakamin Academy

Recommendations:

4. Normal Performance Level Employees:

   - Work Culture Improvement: Conduct surveys and focus groups to understand the specific issues related to toxic culture. Implement initiatives to improve workplace culture, such as diversity and inclusion programs, and employee wellness initiatives.
   - Leadership Development: Invest in leadership training and development programs to ensure managers are equipped to lead effectively and create a supportive work environment.
   - Remote Work Policies: Review and enhance remote work policies to provide more flexibility. Consider hybrid work models that allow employees to work from home part-time.

5. Continuous Feedback and Monitoring:

   - Feedback Loops: Establish regular feedback mechanisms, such as quarterly surveys and pulse checks, to continuously monitor employee satisfaction and engagement.
   - Exit Interviews: Conduct detailed exit interviews to gather insights on why employees are leaving. Use this information to make data-driven improvements to retention strategies.
   - Engagement Initiatives: Implement employee engagement initiatives based on feedback, such as town hall meetings, suggestion boxes, and open-door policies with management.

By addressing the nuanced insights revealed in the visualization, the organization can develop a more holistic and targeted approach to improving employee satisfaction, engagement, and long-term retention across the workforce.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Data Preprocessing

Feature Binning/Mapping

Make the feature binning or mapping to `'Pekerjaan'` and `HiringPlatform`

Train Test Split

Training Set: 75% of the data

Test Set: 25% of the data

Feature Engineering and Selection

The new Feature : `Age, AgeGroup, HiringAssessmentYear, ParticipatedInProject, WasLateLastMonth.`

The feature Selection : `PPScore`

Feature Encoding

Label and One Hot Encoding

Handle Outliers

With Interquartile Range (IQR)

Class Imbalance

SMOTETomek

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Model Evaluation

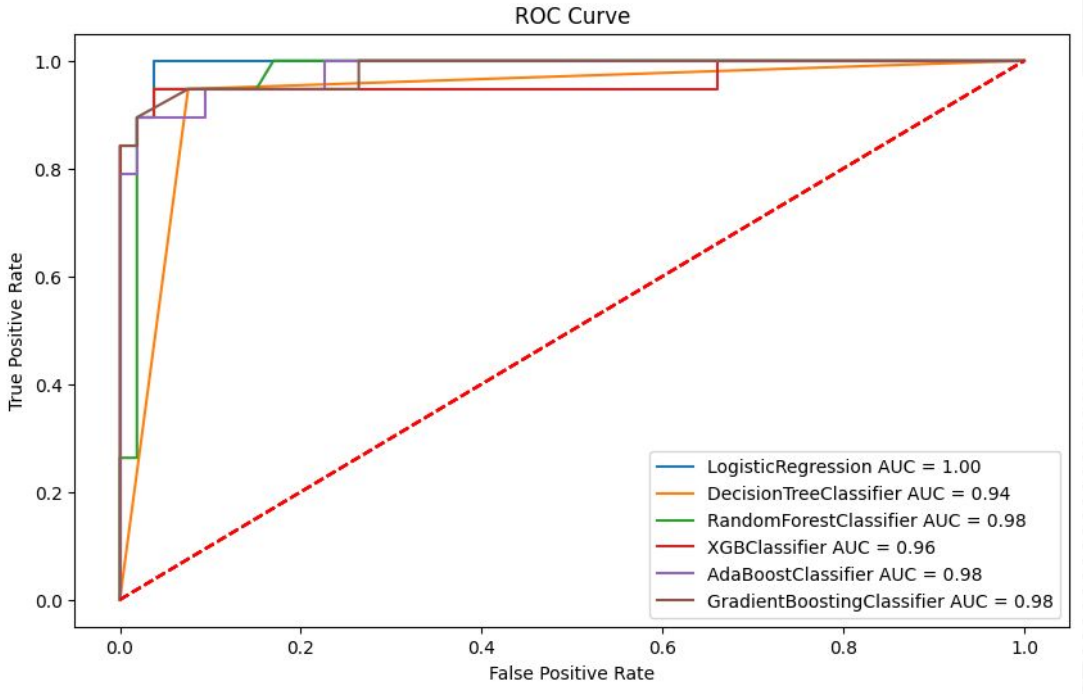
Model	Train AUC	Test AUC	Cross Val AUC	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
Logistic Regression	1.0	0.995035	1.000000	1.0	0.944444	1.0	0.842105	1.0	0.941176	1.0	0.888889
Decision Tree Classifier	1.0	0.935948	0.975714	1.0	0.930556	1.0	0.947368	1.0	0.818182	1.0	0.878049
Random Forest Classifier	1.0	0.977656	0.999490	1.0	0.958333	1.0	0.894737	1.0	0.944444	1.0	0.918919
XGBClassifier	1.0	0.962264	0.997959	1.0	0.958333	1.0	0.894737	1.0	0.944444	1.0	0.918919
AdaBoost Classifier	1.0	0.981132	0.999490	1.0	0.944444	1.0	0.894737	1.0	0.894737	1.0	0.894737
Gradient Boosting Classifier	1.0	0.982622	0.997993	1.0	0.958333	1.0	0.894737	1.0	0.944444	1.0	0.918919

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

ROC Curve



The curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) for various threshold values. The area under the curve (AUC) represents the overall performance of the model. A higher AUC indicates better performance.

All models perform well, with AUC values above 0.9, indicating good accuracy in classifying the target variable. Logistic Regression has the highest AUC (1.0), followed by Random Forest (0.98), Gradient Boosting (0.98), AdaBoost (0.98), and XGBoost (0.96). The Decision Tree performs slightly worse with an AUC of 0.94.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

XGBoost Hyperparameter Tuning

XGBoost Model Performance Before Hyperparameter Tuning (Threshold: 0.22)

Model	Train AUC	Test AUC	Cross Val AUC	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
XGBClassifier	1.0	0.962264	0.997959	0.996479	0.944444	1.0	0.894737	0.993007	0.894737	0.996491	0.894737

XGBoost Model Performance After Hyperparameter Tuning (Threshold: 0.22)

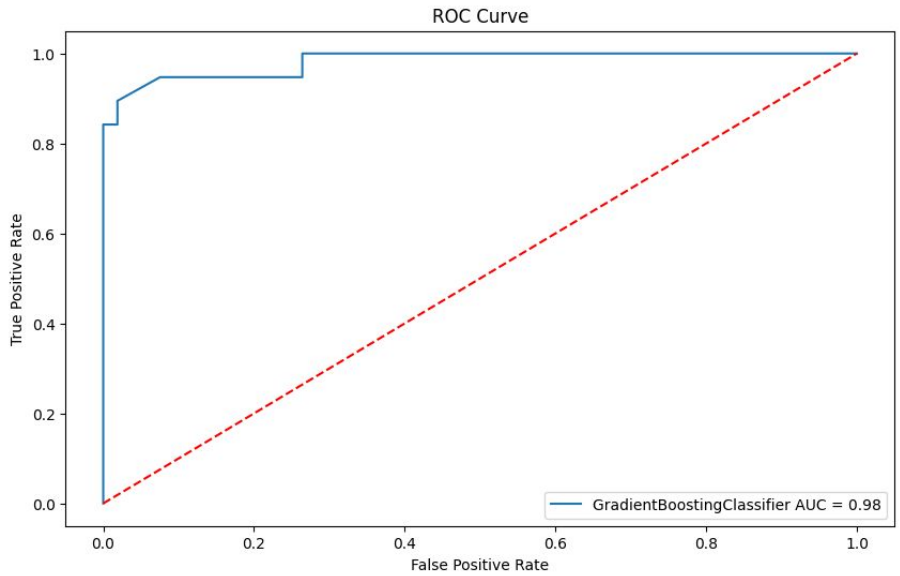
Model	Train AUC	Test AUC	Cross Val AUC	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
XGBClassifier	1.0	0.969215	0.99898	0.992958	0.958333	1.0	0.947368	0.986111	0.9	0.993007	0.923077

Untuk selengkapnya, dapat melihat jupyter notebook disini

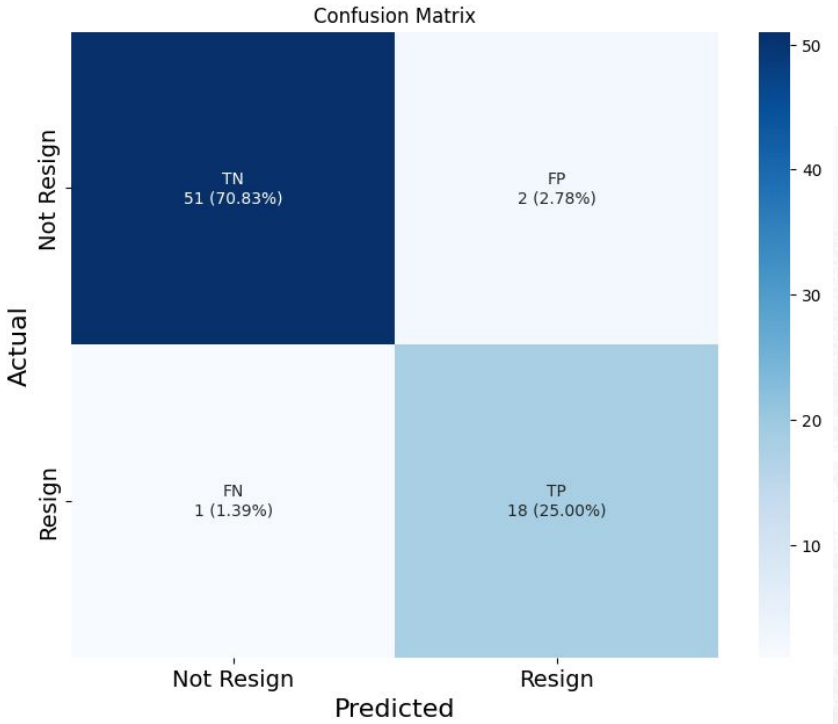
Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

ROC Curve



Confusion Matrix



Untuk selengkapnya, dapat melihat jupyter notebook disini

Insight

The ROC curve plots the true positive (TP) against the false positive(FP) for different classification thresholds. A perfect classifier would have an ROC curve that goes straight up and then straight right, with an area under the curve (AUC) of 1.

The AUC for each classifier is listed in the legend of the ROC curve plot. The higher the AUC, the better the performance of the classifier. In this case, all classifiers have very high AUCs, indicating good performance. The plot shows that all classifiers have very high AUCs, indicating good performance. The best performing classifier is Logistic Regression, with an AUC of 1.0.

The confusion matrix is a table that shows the number of true positives, true negatives, false positives, and false negatives for a classification model. The confusion matrix provides a more detailed view of the classifier's performance, showing how many instances were correctly classified and how many were misclassified. It can be used to assess the accuracy, precision, recall, and F1 score of the classifier.

Untuk selengkapnya, dapat melihat jupyter notebook disini

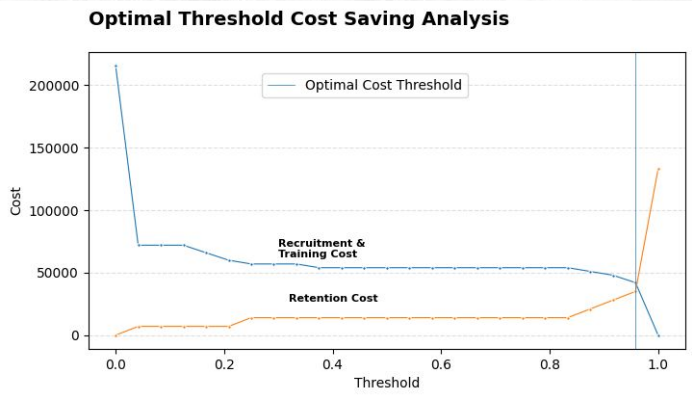
Business Recommendation & Simulation

Business Recommendation & Simulation

Simulation

Define Assumptions:

Cost Category	Cost per Employe
Recruitment Cost	$5000
Training Cost	$2000
Retention Cost	$3000



Optimal Threshold for Cost Savings: 0.96

The graph shows the cost saving analysis for different thresholds. The blue line represents the optimal cost threshold, which is the threshold that minimizes the total cost. The orange line represents the retention cost, which increases as the threshold increases. The green line represents the recruitment and training cost, which decreases as the threshold increases. The optimal threshold is the point where the two lines intersect. In this case, the optimal threshold is approximately 0.95. This means that if the company sets the threshold at 0.95, they will minimize the total cost of recruiting, training, and retaining employees.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

Simulation

Results Table

Metric	*Without Model Simulation*	*With Model Simulation (Default Threshold)*	*With Model Simulation (Optimal Threshold)*
Total Cost Before Model	$288,000.00	-	-
Total Cost After Model	$77,000.00	$68,000.00	$77,000.00
Total Cost Saved	$211,000.00	-	-
Cost Difference (Default vs Optimal)	-	-	-$9,000.00

Cost Savings Analysis with XGBoost Before implementing the XGBoost model, the total cost was 288,000.00. After applying the model, the cost decreased to 77,000.00, resulting in a total cost savings of 211,000.00.

Comparison with Cost Analysis at Default and Optimal Thresholds

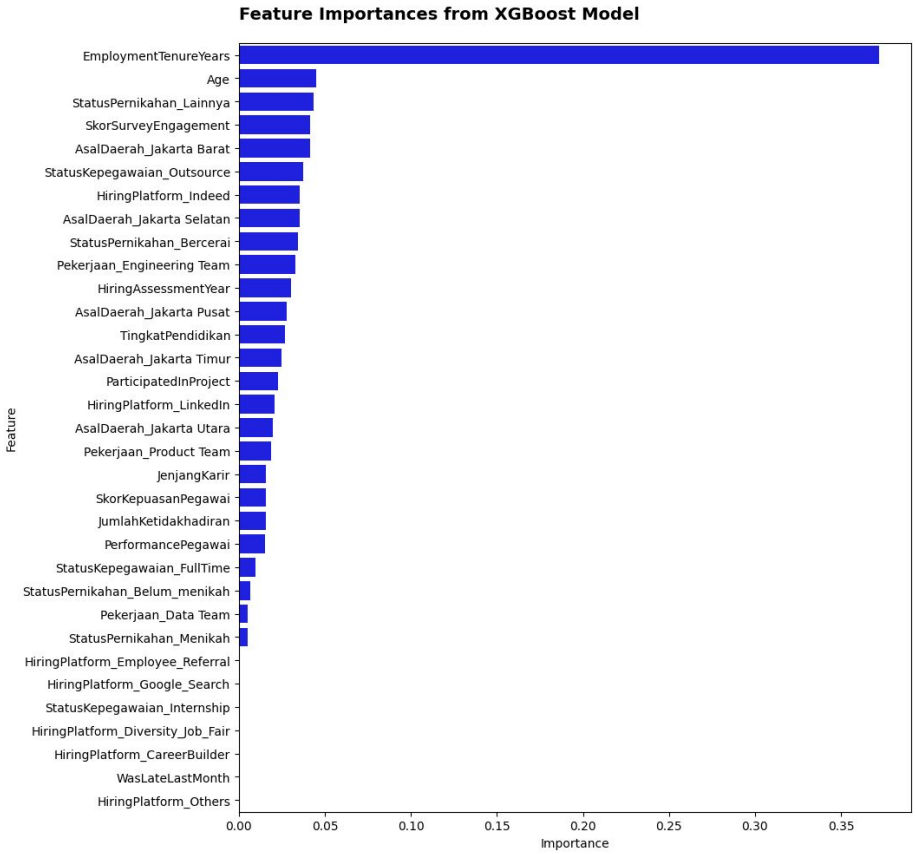
1. Cost at Default Threshold: 68,000.00
2. Cost at Optimal Threshold: 77,000.00
3. Cost Difference (Default vs. Optimal): -9,000.00 (additional cost)

While using the XGBoost model results in significant overall cost savings of 211,000.00, choosing the optimal threshold incurs an additional cost of 9,000.00 compared to the default threshold. This highlights that, despite the overall cost reduction achieved by the model, the optimal threshold may lead to higher costs in certain scenarios.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

Feature Importance From XGBoost Model



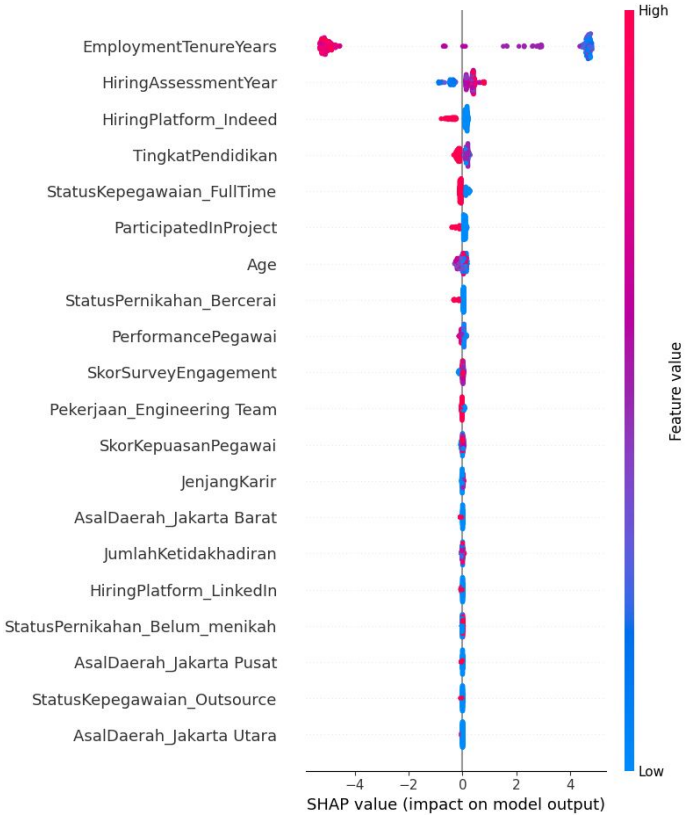
The feature importances from an XGBoost model. The key insights are:

1. The most important feature is `EmploymentTenureYears`, indicating that the employee's tenure years is a significant factor in the model.
2. Other important features include `Age`, `StatusPernikahan_Lainnya`, `SkorSurveyEngagement`, `AsalDaerah_JakartaBarat`, and `StatusKepegawaian_Outsource`.
3. The feature importances range from around 0.05 to 0.30, suggesting none of the individual features are overwhelmingly dominant, but rather the model relies on a combination of several important features.
4. The features cover a variety of employee-related attributes like employment status, location, marital status, and engagement metrics.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

SHAP Value



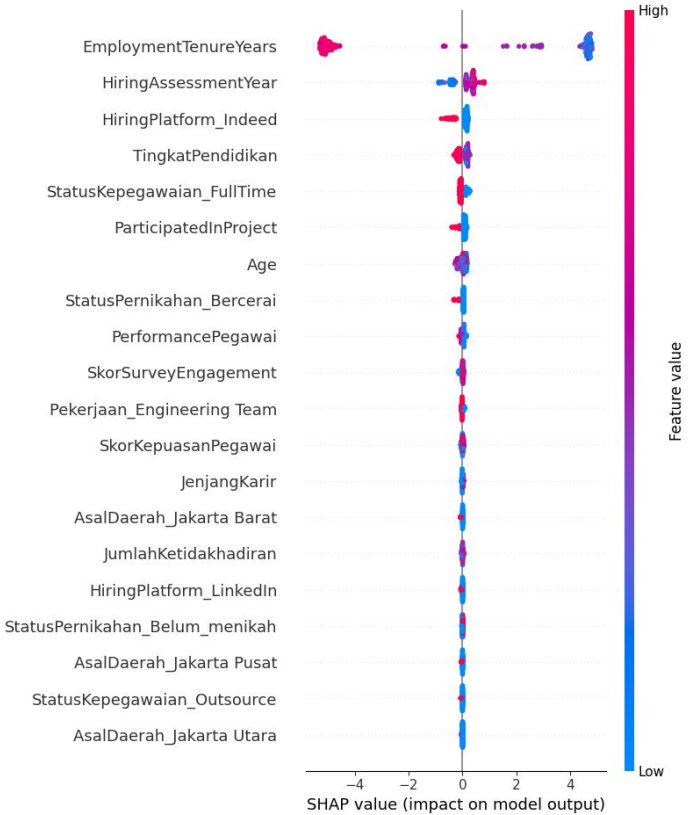
Here's what the SHAP values tell us about each factor's influence on the XGBoost model prediction:

- Employment Tenure (`EmploymentTenureYears`): Employees with shorter tenures are more likely to resign, according to the model. This suggests that building tenure might increase employee retention.
- Hiring Assessment Year (`HiringAssessmentYear`): A higher assessment year (more recent hires) shows a positive impact, indicating newer employees are less likely to resign.
- Hiring Platform (`HiringPlatform_Indeed`, `HiringPlatform_LinkedIn`): Employees hired through Indeed and LinkedIn are predicted to have a lower chance of resigning. This suggests these platforms might attract candidates with a stronger fit for the company.
- Education Level (`TingkatPendidikan`): Higher education levels are associated with a lower chance of resignation, indicating that more educated employees might be more satisfied or engaged with their roles.
- Employment Status (`StatusKepegawaian_FullTime`, `StatusKepegawaian_Outsource`): Full-time employees are less likely to resign, whereas outsourced employees have a higher likelihood of leaving.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

SHAP Value

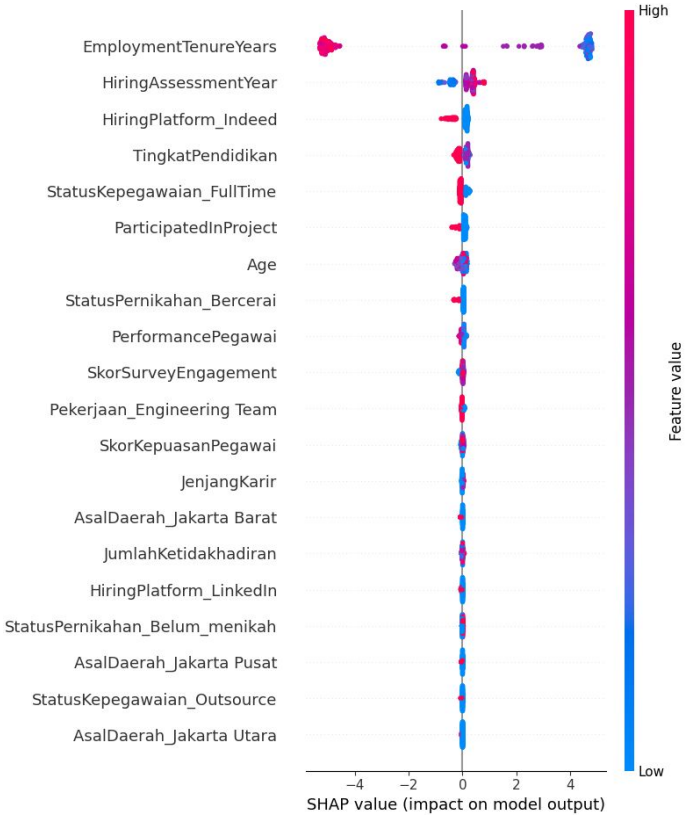


- Project Involvement (`ParticipatedInProject`): Employees who have worked on projects are predicted to have a lower chance of resigning. This suggests that project involvement might increase employee engagement and satisfaction.
- Age: The model predicts a higher chance of resignation for younger employees. This could be due to various factors like career exploration or lack of attachment to the company.
- Marital Status (`StatusPernikahan_Bercerai`, `StatusPernikahan_Belum_menikah`): Being divorced or unmarried has mixed impacts, but overall, these factors are not the most significant predictors.
- Employee Performance (`PerformancePegawai`): The model predicts a higher chance of resignation for high performers. High performers might get better job offers elsewhere, feel under-challenged, or have unmet expectations for advancement.
- Engagement Scores (`SkorSurveyEngagement`): Lower engagement scores predict a higher chance of resignation, highlighting the importance of fostering a positive work environment and addressing employee concerns.
- Employee Satisfaction (`SkorKepuasanPegawai`): Higher satisfaction scores are associated with a lower chance of resignation.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

SHAP Value



- Absenteeism (`JumlahKetidakhadiran`): Interestingly, the model predicts a higher chance of resignation for employees with fewer absences. This could be because highly engaged employees who rarely miss work might be more likely to seek new opportunities if they feel unfulfilled.
- Career Progression (`JenjangKarir`): Employees with better career progression opportunities are less likely to resign, indicating that providing clear career paths can help retain employees.

These insights can help inform strategies to improve employee retention by focusing on factors that contribute to a lower likelihood of resignation.

Untuk selengkapnya, dapat melihat jupyter notebook disini