Rakamin Academy × id/x partners

Prediction Model

ID/X Partners – Data Scientist

Presented by Egi Fermana Putra

Rakamin Academy

egifermn@gmail.com

Egi Fermana Putra

egifermana

Hello! 👋

I'm **Egi Fermana Putra**

A passionate data enthusiast based in Cikarang, Indonesia.

With a Bachelor degree in Nursing from Universitas Medika Suherman, coupled with extensive hands-on experience in data analytics, I'm dedicated to unraveling insights and fostering evidence-based decision-making in diverse domains.

Proficient of statistical methods such as chi-square, t-test, and linear regression, etc. Equipped with hands-on experience in Python libraries, R Studio, SQL, and SPSS Statistics, adeptly analyzing data to extract meaningful insights.

Skilled in crafting compelling data visualizations and dashboard using Tableau, Looker Studio, and Cognos Analytics. Knowing Microsoft Excel and Google Sheets. Committed to delivering impactful results through data-driven strategies.

Courses and Certification ✅

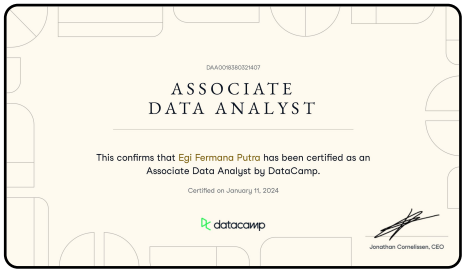
Rakamin Academy

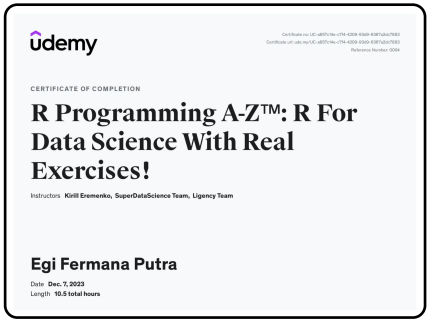
Data Analyst Associate | https://www.datacamp.com/certificate/DAA0018380321407 — **January, 2024**

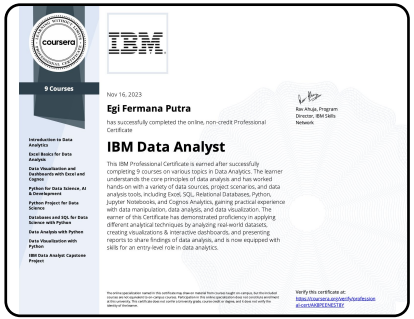
IBM Data Science | https://www.coursera.org/account/accomplishments/specialization/JRPWTG5832ME — **December, 2023**

R Programming A-Z | https://www.udemy.com/certificate/UC-a857c14e-c7f4-4209-93d9-6367a2dc7883 — **December, 2023**

IBM Data Analyst | https://www.coursera.org/account/accomplishments/professional-cert/AK8PEENEST8Y — **November, 2023**







About Company 💊

Rakamin Academy

ID/X Partners (PT IDX Consulting) was established in 2002 and has served companies across Asia and Australia, particularly in financial services, telecommunications, manufacturing, and retail industries. Specializing in leveraging data analytics and decisioning solutions (DAD) combined with risk management and integrated marketing discipline, ID/X Partners provides consultancy services to help clients optimize portfolio profitability and business processes. The comprehensive consultancy services and technology solutions offered by ID/X Partners make it a one-stop service provider.

id/x partners

Project Portfolio 😎

Rakamin Academy

As a Data Scientist at ID/X Partners, the project involves collaboration with a lending company (multifinance) client, where the objective is to enhance the accuracy of credit risk assessment and management. This improvement is vital for optimizing the client's business decisions and reducing potential losses.

The task at hand is to develop a machine learning model capable of predicting credit risk based on the provided dataset, encompassing both approved and rejected loan data. The development process includes several stages, starting with Data Understanding, Exploratory Data Analysis (EDA), Data Preparation, Data Modeling, and Evaluation.

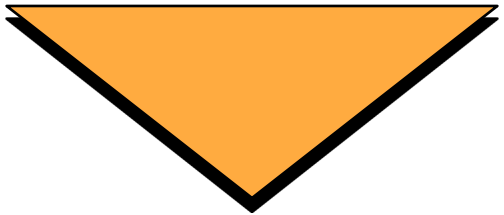
Contributing to this project will enable the client to make informed decisions regarding credit risk, ultimately enhancing their business performance and minimizing financial risks.

GitHub repository here!

The Challenges 🎯

Rakamin Academy

Challenge #1	Challenge #2	Challenge #3
Data Understanding, Exploratory Data Analysis	Data Preparation, Data Modelling, Evaluation	Machine Learning Algorithm



Objective

The objective of this project is to develop a machine learning model which can predict credit risk based on dataset provided, which includes loan data approved and rejected.

Tools and Datasets ⚙️

Rakamin Academy

Tools

Pyhton 

Jupyter Notebook 

Datasets

Source: Loan Dataset (*Link*).

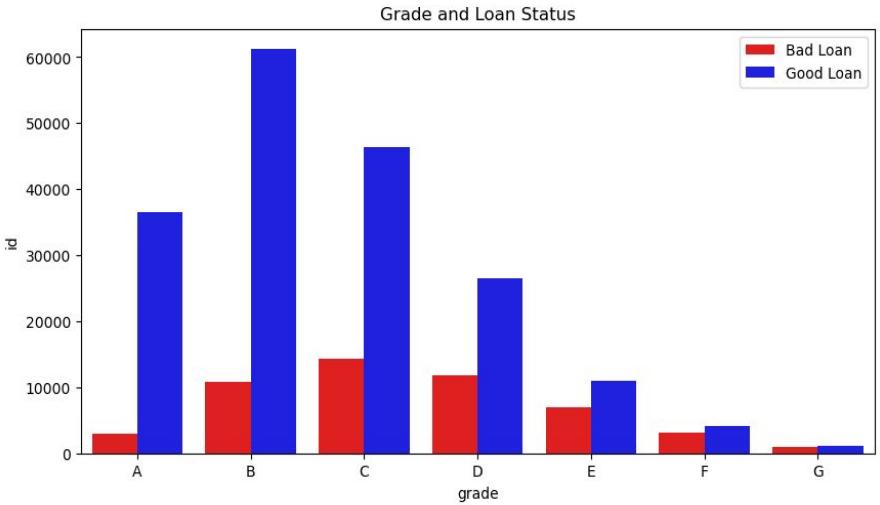
Data Understanding 🔥

Rakamin Academy

Our dataset comprises **466,285** entries spread across **58 columns**. These columns include **numeric (29)**, **integer (7)**, and **categorical (22)** data types. We've identified 1**7 columns with missing values**. Notably, four columns have over 50% missing data, including **'mths_since_last_record'**, **'mths_since_last_major_derog'**, **'desc'**, and **'mths_since_last_delinq'**. Addressing these gaps is crucial for reliable analysis. Additionally, **'next_payment_d'** shows **40-50% missingness**, impacting payment behavior insights. While **12 columns have 1-20% missing values**, such as **'tot_cur_bal'** and **'emp_title'**. Despite missing values, there are no duplicate entries, ensuring data integrity.

Exploratory Data Analysis 🔥

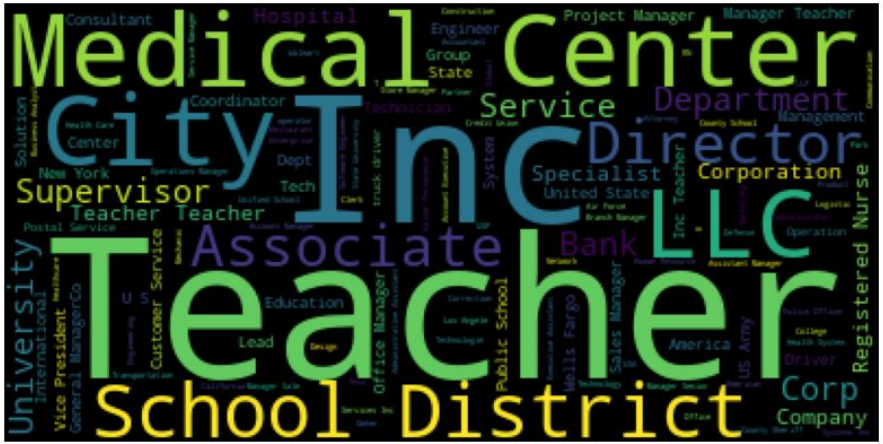
Rakamin Academy



- **Loan Status by Grade:** The chart divides loan status into two categories: Bad Loan and Good Loan. It shows the percentage of borrowers with each loan status for each grade level (A through G).
- **Distribution of Grades:** The percentage of borrowers seems to be fairly evenly distributed across all grades, with slightly higher percentages for grades B, C, and D.
- **Loan Performance by Grade:** Interestingly, the percentage of borrowers with bad loans appears to be higher than the percentage of borrowers with good loans for all grades. This suggests that a higher proportion of borrowers across all grade levels end up with bad loans.
- **Least Risky Grades:** Grade A appears to have the lowest percentage of bad loans, followed by Grade B. This could indicate that borrowers with higher grades tend to have a better credit history or are more likely to manage their loans successfully.
- **Most Risky Grades:** Conversely, Grade G appears to have the highest percentage of bad loans, followed by Grade F. This could suggest that borrowers with lower grades may have a weaker credit history or face challenges repaying their loans.

Exploratory Data Analysis 🔥

Rakamin Academy



- **Focus on Borrower Attributes:** The word cloud emphasizes terms like "medical center," "teacher," "city director," and "annual income," suggesting a focus on the borrower's background and financial stability when evaluating loan applications.
- **Potential Borrower Demographics:** The inclusion of terms like "teacher" and "city director" might indicate the loan provider targets a specific demographic or profession.
- **Creditworthiness Factors:** Words like "credit," "income," "employment," and "dti" (debt-to-income) highlight the importance of a borrower's financial history and ability to repay the loan.
- **Loan Properties:** The presence of terms like "loan amount" and "term" indicates these are likely factors considered during the loan application process.

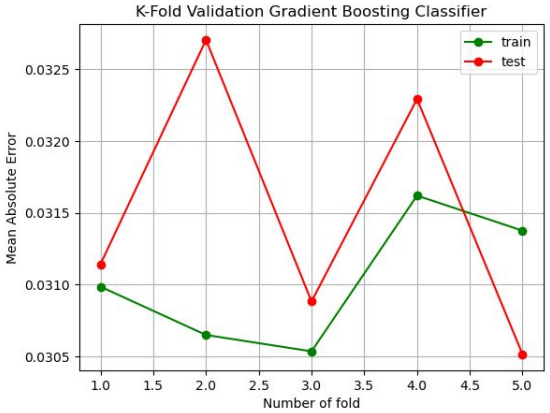
Data Preparation

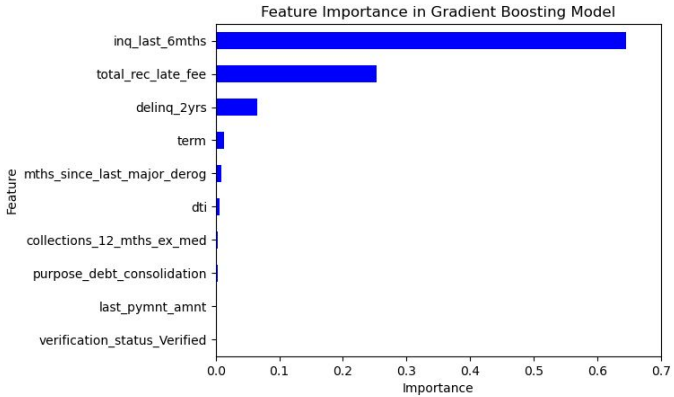
Rakamin Academy

Missing Values	First, we remove columns with a significant number of missing values or those not relevant to the analysis, which helps reduce noise and focus on important data. Second, for remaining missing values, we use imputation, filling them in with estimated values. Imputation methods include using statistical measures like the mean, median, or mode for numerical data, or the most common category for categorical data. This ensures our dataset remains informative and complete, avoiding biases in our analysis.
Feature Engineering	Feature engineering involves transforming categorical variables into a format interpretable by machine learning algorithms, which can be achieved through techniques like one-hot encoding, label encoding, or target encoding. Additionally, encoding converts categorical variables into numerical representations, offering methods such as one-hot encoding, label encoding, and target encoding. Log transformation normalizes skewed or non-normally distributed data, while standardization scales numerical features for easier comparison across different units and magnitudes.
Feature Selection	Correlation analysis assesses the connection between variables by calculating correlation coefficients, indicating the strength and direction of linear relationships. High values signify strong correlations, while low values suggest weak or no relationships. Feature selection using correlation aids in identifying key features for analysis and mitigating multicollinearity issues. Eliminating highly correlated features can enhance machine learning model performance and interpretability.

Data Modelling

Rakamin Academy





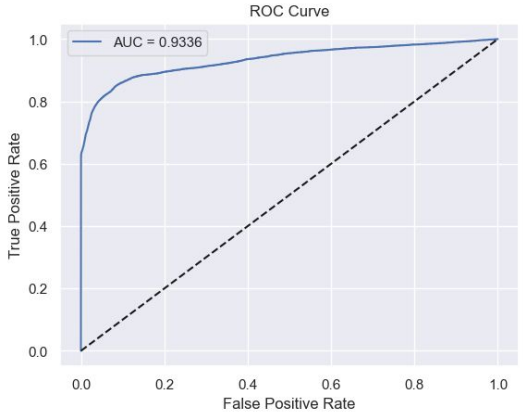
Model Checking:

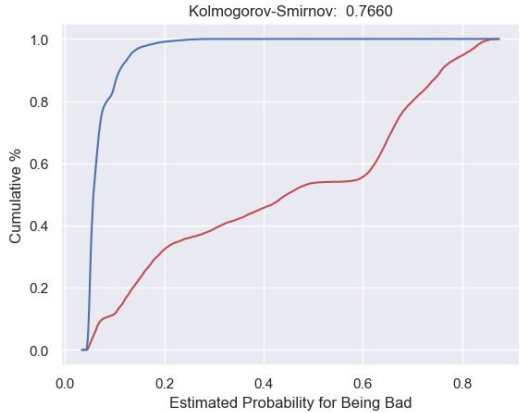
Utilizing K-Fold Cross Validation: The code snippet showcases the K-Fold Cross Validation technique for model evaluation. It splits the dataset into training and testing sets, trains the model, predicts on the test set, and calculates the Mean Absolute Error (MAE) for both training and testing data across multiple folds. The resulting MAE scores are plotted against the number of folds, providing insights into the model's performance and potential overfitting or underfitting.

Model Interpretation:

Feature Importance Analysis: The provided function computes the feature importance of the Gradient Boosting Classifier model. By aggregating feature importance scores from individual estimators within the model, it offers insights into which features contribute the most to predictive performance. The resulting feature importance scores are visualized through a horizontal bar plot, highlighting the top 10 influential features in the model. This analysis aids in understanding the underlying mechanisms driving the model's predictions and identifying key factors influencing the target variable.

Evaluation





Area Under the Curve (AUC):

AUC is a common metric used to measure the performance of credit risk models. It quantifies the model's ability to distinguish between positive and negative cases, with higher values indicating better discrimination. The Receiver Operating Characteristic (ROC) curve visually represents the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity). For the current model, the AUC value is calculated to be 0.9336, indicating strong predictive performance.

Kolmogorov-Smirnov (KS):

KS is another crucial metric in credit risk modeling, measuring the maximum difference between the cumulative distributions of positive and negative cases. A higher KS value signifies greater model discriminatory power. The KS plot illustrates the cumulative percentage of bad and good cases against the estimated probability for being bad. In this model, the KS value is computed as 0.7660, demonstrating substantial discriminatory ability.

Interpretation:

The obtained model exhibits robust performance, with an AUC above 0.7 and a KS above 0.3, which are commonly regarded as thresholds for good performance in credit risk modeling. These metrics validate the model's effectiveness in accurately predicting credit risk, providing confidence in its utility for decision-making purposes.

Summary and Feedback 📋

Rakamin Academy

Summary Insight

- If higher interpretability is desired, consider creating a Credit Scorecard using the Logistic Regression algorithm with approaches such as Feature Selection using Information Value and Feature Engineering using Weight of Evidence.
- If interpretability is not a high priority, consider trying other machine learning algorithms such as Boosting.
- Perform hyperparameter tuning.
- Ensure that the model created is not overfitting. This can be done by comparing the performance of the model when predicting against training data and when predicting against testing data.
- Generally, the more appropriate step is to perform Train-Test Split before feature transformation such as encoding or scaling. However, for simplicity reasons, this example does the opposite because typically the difference in performance is not significantly different.

Thank You

Rakamin Academy × id/x partners