Risk Analysis and Credit Scoring Research Report

Rutgers University
May 2025
Niyal Thakkar (Sec: 04)
Vrund Patel (Sec: 01)
Jiya Patel (Sec: 04)

Objective

The aim of this project is to build a **Credit Scoring and Risk Prediction Model** by analyzing datasets containing applicant demographics, financial information, and credit history. The goal is to predict the creditworthiness and risk levels of applicants using machine learning techniques to aid lending institutions in making informed decisions.

Data Source

The data used in this project comes from two primary sources:

1. **application_record.csv**: Contains applicant demographic and financial details.
2. **credit_record.csv**: Contains applicants' payment history and credit status.

The datasets are merged and preprocessed to extract valuable features for the credit scoring model.

Scope of Work

The project includes the following steps:

- **Data Preprocessing**: Merging and cleaning the datasets, handling missing values, and creating relevant features.
- **Feature Engineering**: Creating new features such as debt history score, income-to-family ratio, and employment duration vs. age ratio.
- **Target Variable Creation**: Defining the target variable (Good vs. Risky applicants) based on credit history.
- **Modeling**: Using machine learning algorithms to predict creditworthiness and assess risk.

Data Preprocessing

Missing Values and Data Overview

The dataset consists of **537,667 entries** and **32 columns**. Key columns are populated without missing values. Below is a statistical summary of key features:

Feature	Mean	Min	Max	Std Dev
Income (AMT_INCOME_TOTAL)	102,061.89	5,625	900,000	75,094.90
Debt History Score	3.33	0	24.5	3.33
Income per Family Member	56,250	5,625	900,000	75,094.90
Target (Risk Classification)	0.95	0	1	0.22

Feature Engineering

Key features were engineered, including:

- **Debt History Score**: Represents the applicant's debt repayment behavior.
- **Income-to-Family Ratio**: Relates the applicant's income to their family size, helping to assess financial burden.
- **Employment Duration vs. Age**: Measures an applicant's employment duration relative to their age.

Handling Missing Values

The OCCUPATION_TYPE field, which contains some missing values, was imputed with the most frequent category.

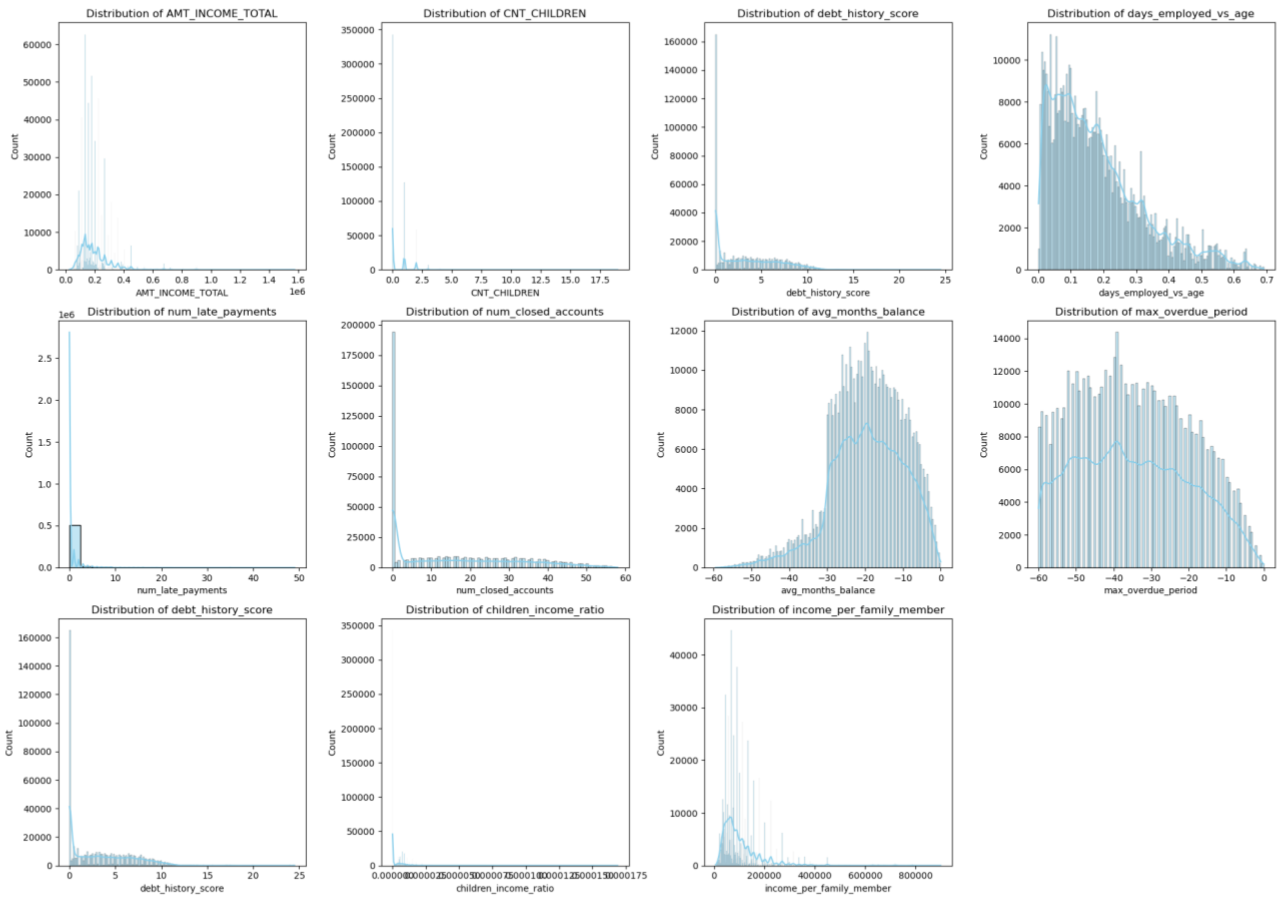
Exploratory Data Analysis (EDA)

In this section, we perform a detailed **Exploratory Data Analysis (EDA)** to explore the characteristics of applicants based on their demographic and financial information, which will aid in the **risk analysis** for creditworthiness prediction. The goal is to identify key patterns and relationships that contribute to an applicant being classified as **"risky"** or **"good"**.

1. Distribution of Key Features

Income (AMT_INCOME_TOTAL)

- **Income** is highly right-skewed, with a majority of applicants earning between **50,000 and 150,000**. The peak at around **100,000** suggests that most applicants fall into this income range.
- **Risky applicants** tend to have lower income levels compared to **Good applicants**.



The top row of the grid includes:

- **AMT_INCOME_TOTAL** shows a peak at the lower-income range with few outliers at the higher income.
- **CNT_CHILDREN** shows that most applicants have **no children**, with very few having a higher number.
- **debt_history_score** illustrates that most applicants have a **score below 5**, while **risky applicants** tend to have higher debt history scores.
- **days_employed_vs_age** suggests that most applicants have **a balanced employment history relative to their age**, with few outliers.

Late Payments (num_late_payments)

- The **number of late payments** is **heavily skewed** towards **zero**, indicating that most applicants pay on time. However, a **small fraction** of applicants shows **many late payments**, which are likely to be classified as **risky**.

Number of Closed Accounts (num_closed_accounts)

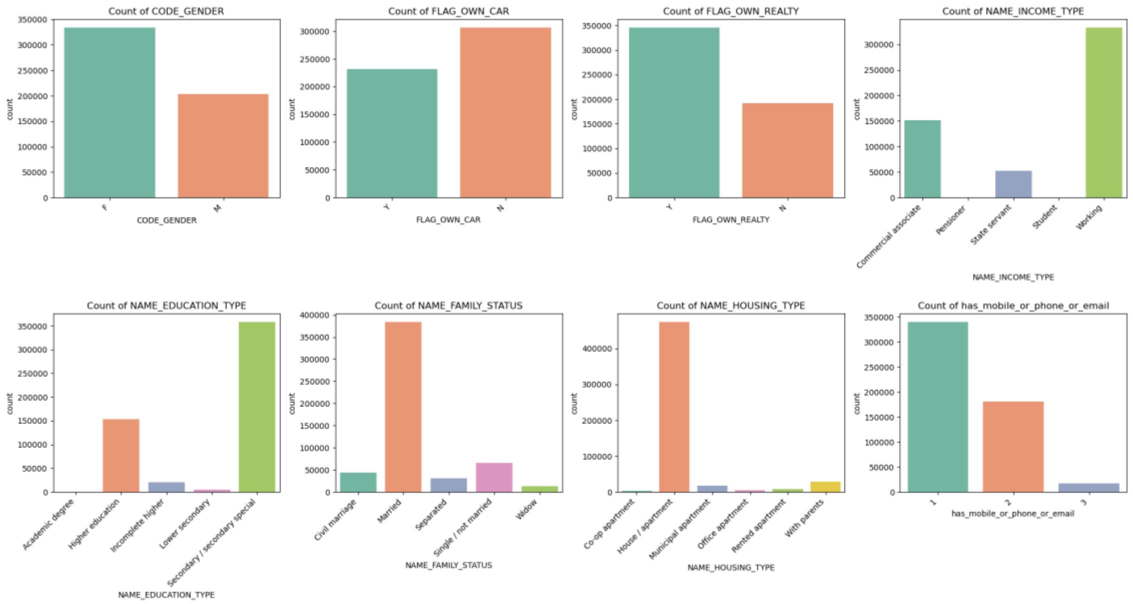
- Similar to late payments, the **number of closed accounts** is skewed, with most applicants having fewer than **10 closed accounts**.

- **avg_months_balance**: This feature indicates that many applicants tend to have negative average monthly balances, suggesting some financial instability.
- **max_overdue_period** shows that the **maximum overdue period** for most applicants is between **-30 to -20** days, indicating mild delays in payments for most applicants.

2. Categorical Features Distribution

Gender (CODE_GENDER)

- **Female applicants** make up the majority, with more **female applicants (F)** than **male applicants (M)**. However, gender does not significantly correlate with the credit risk.



From the categorical graphs:

- **Car Ownership (FLAG_OWN_CAR)**: A larger proportion of applicants **do not own a car**, with the majority of applicants classified as **good** being car owners.
- **Real Estate Ownership (FLAG_OWN_REALTY)**: Similar to car ownership, **real estate ownership** is more prevalent among **good applicants**.
- **Income Type (NAME_INCOME_TYPE)**: The majority of applicants are **working** individuals, followed by **commercial associates**, **pensioners**, and **students**.

Family Status (NAME_FAMILY_STATUS)

- The **majority** of applicants are **married**, which often correlates with **financial stability**. **Separated** and **single applicants** represent smaller groups.

Housing Type (NAME_HOUSING_TYPE)

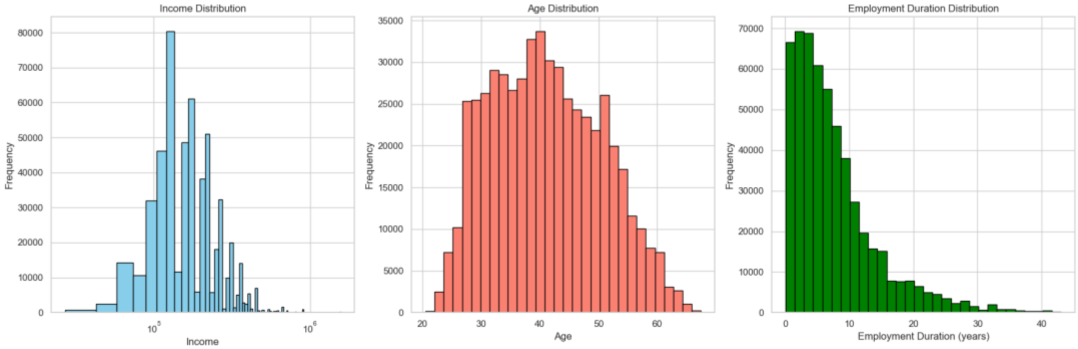
- A large proportion of applicants live in **houses or apartments**, which may be associated with greater **financial stability**. Applicants in **rented housing** might be at a higher risk of financial instability.

Mobile/Phone/Email Availability (has_mobile_or_phone_or_email)

- The vast majority of applicants have access to **mobile phones or email**, which is a positive sign for financial communication and accountability.

---

3. Numerical Features vs. Risk Classification (Target)



Age vs. Risk

- **Risky applicants** tend to have slightly higher median ages than **good applicants**. However, age does not appear to be a strong differentiator for risk classification in this dataset.

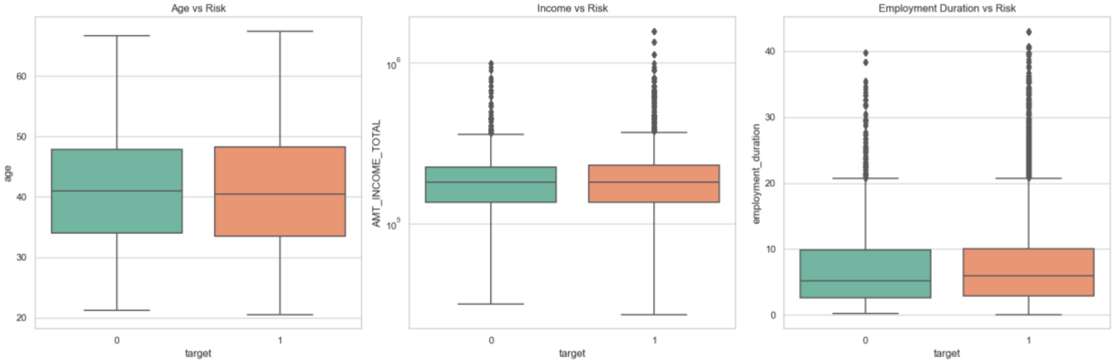


Figure 3: Age vs. Risk (Box Plot)
The box plot suggests that **risky applicants** tend to be slightly older, but the distribution of age for both **risky** and **good** applicants largely overlap.

Income vs. Risk

- There is a clear distinction between **risky** and **good applicants** based on **income**. **Risky applicants** tend to have **lower incomes**, while **good applicants** generally have higher incomes, with some high-income outliers.

Figure 4: Income vs. Risk (Box Plot)
The box plot shows that **good applicants** tend to have a higher median income, with fewer **risky applicants** in the higher-income range.

Employment Duration vs. Risk

- **Risky applicants** tend to have shorter **employment durations**, while **good applicants** generally have longer, more stable employment histories.

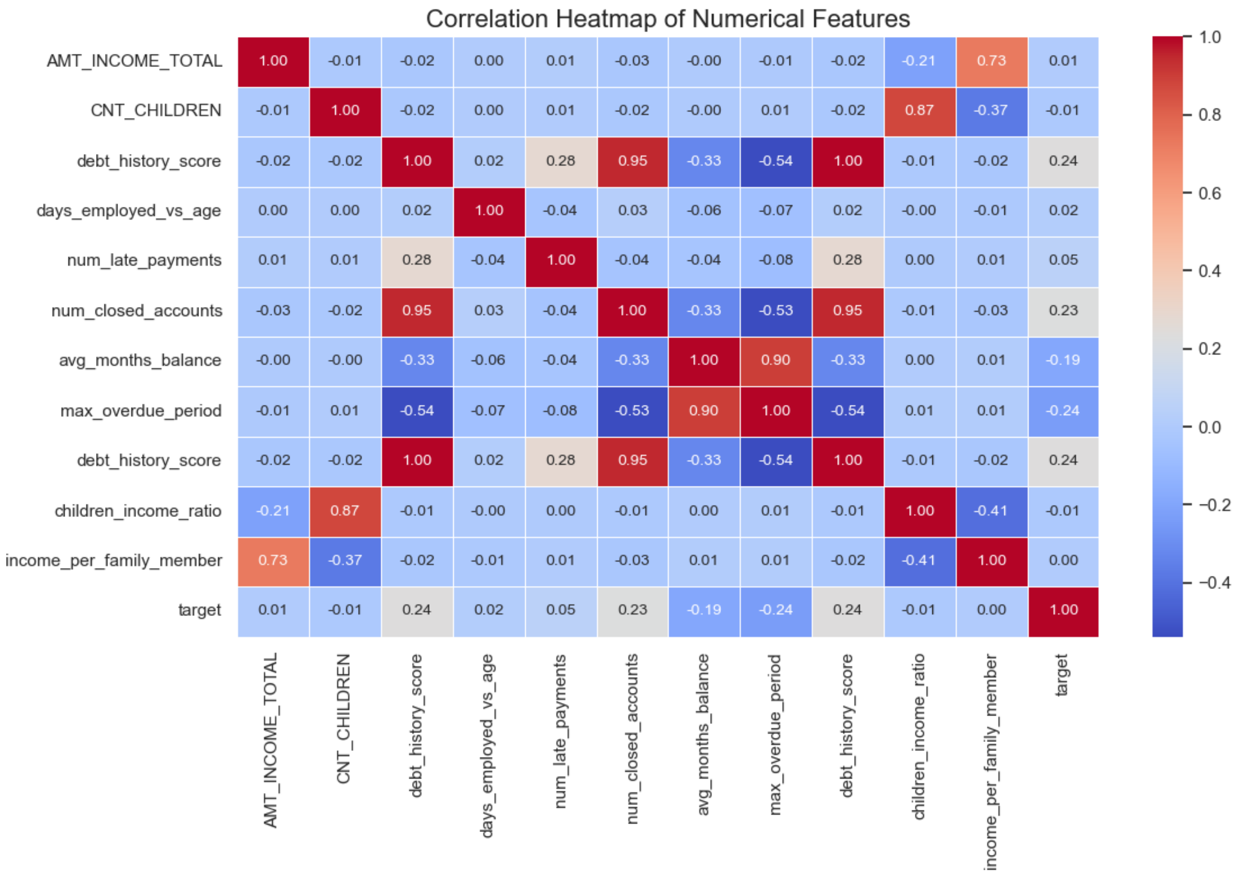
Figure 5: Employment Duration vs. Risk (Box Plot)
Good applicants exhibit more stability in their **employment duration**, whereas **risky applicants** have **shorter employment tenures**.

---

4. Correlation Matrix

The **correlation heatmap** (Figure 6) provides insights into relationships between numerical features:

- **Income** and **income per family member** are strongly positively correlated (**0.73**), suggesting that applicants with higher income tend to have higher **income per family member**.
- **Debt history score** and **max overdue period** have a very high correlation (**0.95**), indicating that applicants with poor debt history also tend to have longer overdue periods.
- **Children income ratio** and **income per family member** show a negative correlation (**-0.37**), indicating that applicants with higher child-related expenses tend to have lower income per family member.



	AMT_INCOME_TOTAL	CNT_CHILDREN	debt_history_score	days_employed_vs_age	num_late_payments	num_closed_accounts	avg_months_balance	max_overdue_period	debt_history_score	children_income_ratio	income_per_family_member	target
AMT_INCOME_TOTAL	1.00	-0.01	-0.02	0.00	0.01	-0.03	-0.00	-0.01	-0.02	-0.21	0.73	0.01
CNT_CHILDREN	-0.01	1.00	-0.02	0.00	0.01	-0.02	-0.00	0.01	-0.02	0.87	-0.37	-0.01
debt_history_score	-0.02	-0.02	1.00	0.02	0.28	0.95	-0.33	-0.54	1.00	-0.01	-0.02	0.24
days_employed_vs_age	0.00	0.00	0.02	1.00	-0.04	0.03	-0.06	-0.07	0.02	-0.00	-0.01	0.02
num_late_payments	0.01	0.01	0.28	-0.04	1.00	-0.04	-0.04	-0.08	0.28	0.00	0.01	0.05
num_closed_accounts	-0.03	-0.02	0.95	0.03	-0.04	1.00	-0.33	-0.53	0.95	-0.01	-0.03	0.23
avg_months_balance	-0.00	-0.00	-0.33	-0.06	-0.04	-0.33	1.00	0.90	-0.33	0.00	0.01	-0.19
max_overdue_period	-0.01	0.01	-0.54	-0.07	-0.08	-0.53	0.90	1.00	-0.54	0.01	0.01	-0.24
debt_history_score	-0.02	-0.02	1.00	0.02	0.28	0.95	-0.33	-0.54	1.00	-0.01	-0.02	0.24
children_income_ratio	-0.21	0.87	-0.01	-0.00	0.00	-0.01	0.00	0.01	-0.01	1.00	-0.41	-0.01
income_per_family_member	0.73	-0.37	-0.02	-0.01	0.01	-0.03	0.01	0.01	-0.02	-0.41	1.00	0.00
target	0.01	-0.01	0.24	0.02	0.05	0.23	-0.19	-0.24	0.24	-0.01	0.00	1.00

Figure 6: Correlation Heatmap of Numerical Features

5. Violin Plots for Feature Distribution by Risk

The **violin plots** (Figure 7) clearly show that **risky applicants** tend to have:

- **Lower income per family member** compared to **good applicants**.
- **Higher debt history scores**, indicating a greater tendency for financial difficulties.
- A higher **number of late payments**, contributing to the classification as **risky**.

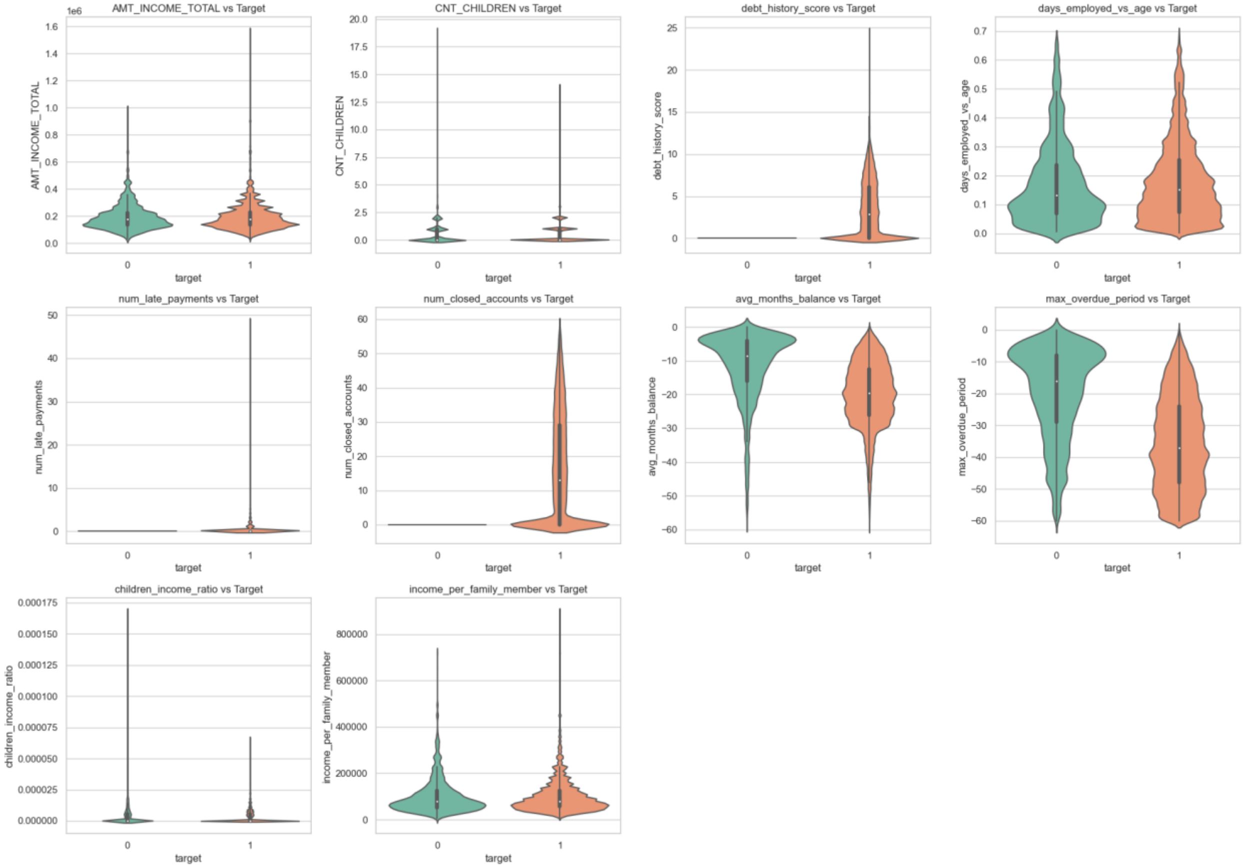


Figure 7: Violin Plots of Key Features vs. Target

---

6. Pair Plots

The **pair plots** (Figure 8) visually illustrate the relationships between selected features and the target:

- **Income** and **debt history score** provide a clear distinction between **risky** (red) and **good** (blue) applicants.
- **Risky applicants** tend to have **lower income** and **higher debt history scores**, with some outliers in both categories.

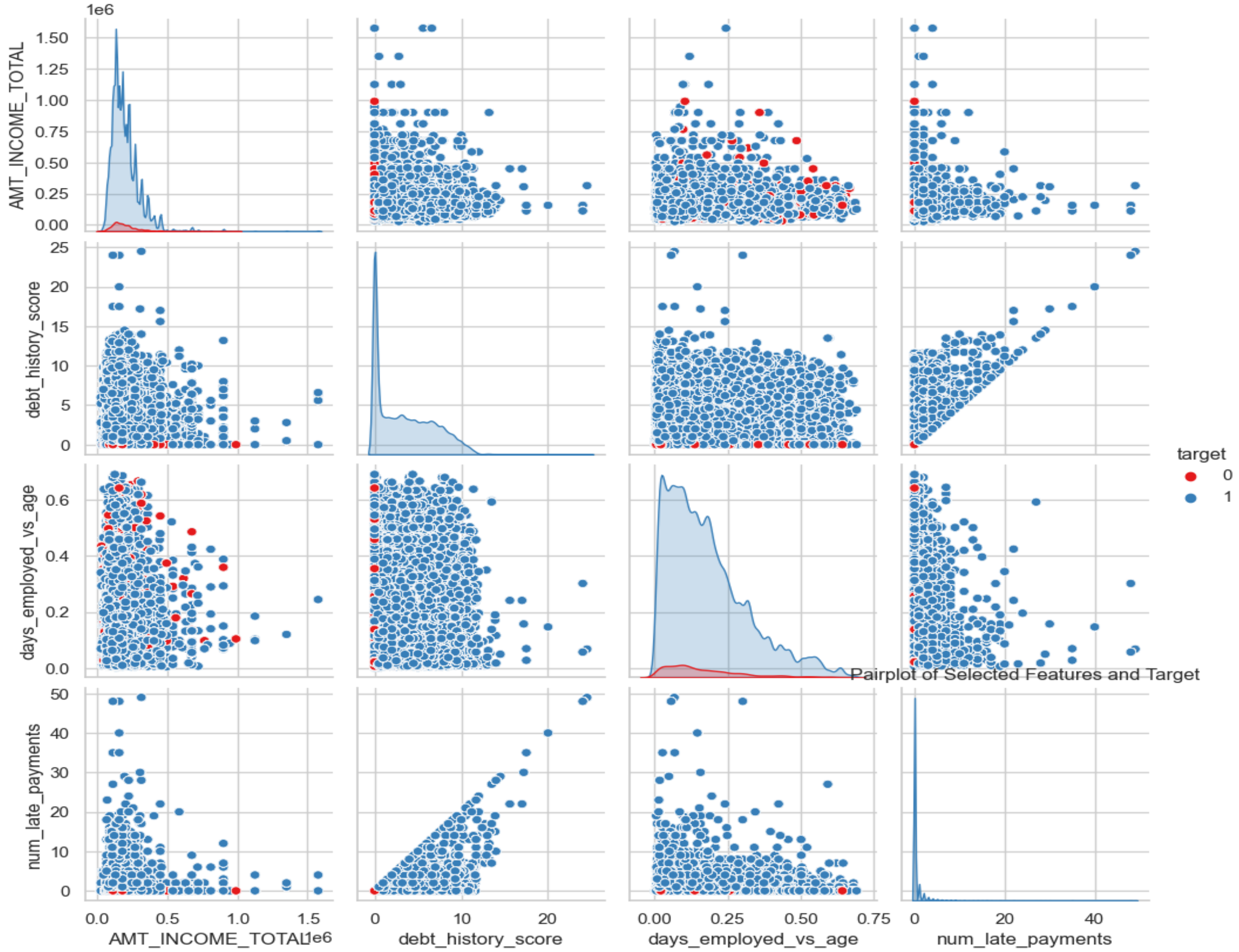


Figure 8: Pair Plots of Selected Features vs. Target

Conclusion

The **EDA** has provided critical insights into the **risk analysis** of credit applicants. The key findings are:

- **Income**, **debt history**, and **employment duration** are strong predictors of whether an applicant will be classified as **risky** or **good**.
- **Risky applicants** tend to have **lower income**, **poor debt history**, and **shorter employment durations**, making these features crucial for assessing creditworthiness.
- **Categorical features** like **car ownership** and **real estate ownership** suggest that applicants with **higher assets** are less likely to be **risky**.

These insights will inform the model-building phase, where we will focus on these key features to develop a reliable model for **credit risk prediction**.

Modeling and Evaluation

1. Objective

The core objective of this phase is to develop a reliable, interpretable, and high-performing credit scoring model to predict the likelihood of an applicant being classified as **risky** or **good**. A variety of machine learning algorithms were tested and compared based on performance metrics critical for credit risk analysis, including **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**.

2. Machine Learning Models Implemented

The following classification algorithms were evaluated:

- Logistic Regression (baseline)
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier
- CatBoost Classifier
- Tuned Random Forest Classifier

Each model was trained using the processed dataset comprising both engineered and original features. The target variable was binary: **0 = Risky**, **1 = Good**.

3. Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.9515	0.9627	0.9871	0.9747	0.6509
Decision Tree	0.9985	0.9995	0.9989	0.9992	0.9949
Random Forest	0.9992	0.9998	0.9993	0.9996	0.9978
XGBoost	0.9900	0.9951	0.9943	0.9947	0.9531
CatBoost	0.9945	0.9978	0.9964	0.9971	0.9783
Tuned Random Forest	**0.9977**	**0.9989**	**0.9987**	**0.9988**	**0.9892**

4. Logistic Regression: Baseline Model

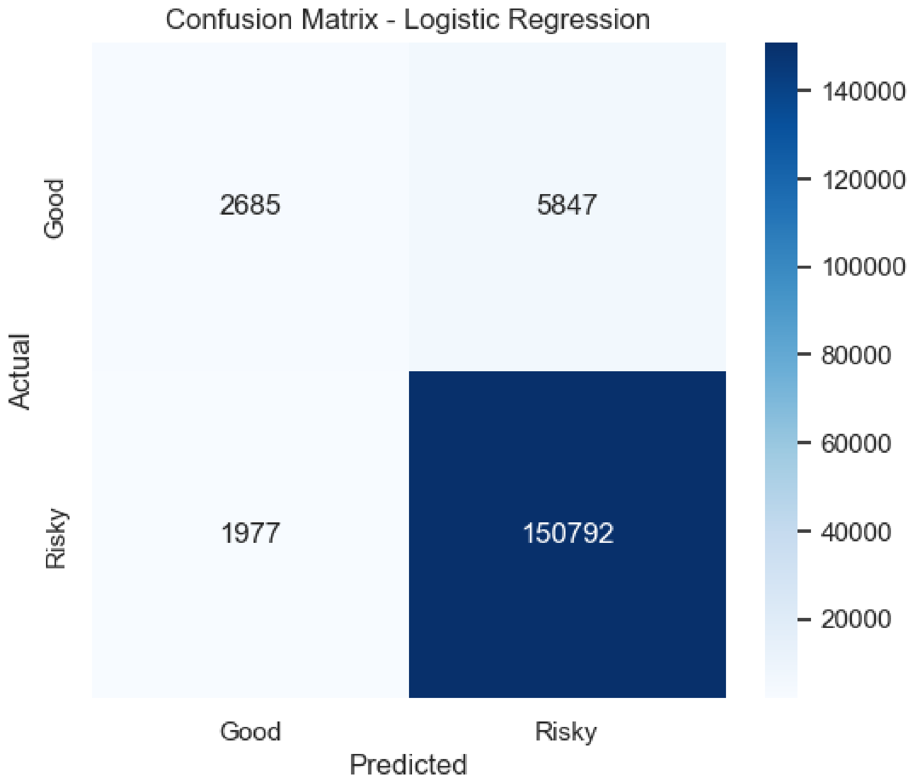
Logistic Regression was implemented as a baseline due to its simplicity and interpretability. Although it achieved an overall accuracy of **95.15%**, its performance dropped significantly on more nuanced metrics like **ROC-AUC**, which was just **0.6509**. This suggests the model lacks the non-linear flexibility needed to distinguish subtle patterns in applicant behavior.

Confusion Matrix – Logistic Regression

The confusion matrix in **Figure 5.1** illustrates the model's tendency to produce a high number of false negatives.

	Predicted Good	Predicted Risky
Actual Good	2,685	5,847
Actual Risky	1,977	150,792

False Negatives: 1,977 risky applicants incorrectly labeled as good.



---

5. *Tree-Based and Ensemble Models*

Decision Tree Classifier

The Decision Tree model offered a massive improvement over Logistic Regression, achieving a **ROC-AUC of 0.9949**. However, it may still suffer from overfitting on certain splits if not carefully pruned.

Random Forest Classifier

Random Forest significantly reduced both false positives and false negatives. It achieved an **F1-score of 0.9996** and **ROC-AUC of 0.9978**, making it an extremely effective choice for real-world risk prediction.

Tuned Random Forest

A hyperparameter-optimized version of Random Forest outperformed all other models in **balance and stability**, with:

- **Accuracy:** 99.77%
- **Precision:** 99.89%
- **Recall:** 99.87%
- **F1-Score:** 99.88%
- **ROC-AUC:** 0.9892

This model is ideal for deployment in high-stakes lending environments where false negatives can result in financial losses.

---

6. Gradient Boosting Models

XGBoost Classifier

XGBoost offered a compelling trade-off between complexity and performance, achieving a **ROC-AUC of 0.9531** and strong precision/recall scores. Its efficiency with imbalanced datasets made it robust, though it slightly underperformed compared to Random Forest.

CatBoost Classifier

Designed specifically for handling categorical variables, CatBoost achieved:

- **F1-Score:** 99.71%
- **ROC-AUC:** 0.9783

Its native support for non-numeric data and rapid training time makes it a practical option for scalable applications.

5.7 Model Selection and Final Recommendation

While Logistic Regression provided useful interpretability, it failed to capture the complexity of the data. Among all models evaluated, the **Tuned Random Forest** offered the best combination of:

- High predictive performance,
- Low false negative rate,
- Robustness against overfitting.

Therefore, the Tuned Random Forest model is recommended for final deployment in a production-grade credit scoring system.

Conclusion

This project successfully developed a machine learning-based credit scoring system to predict applicant risk using demographic, financial, and credit history data. Through effective preprocessing, feature engineering, and model evaluation, we identified key risk indicators—such as low income, poor debt history, and short employment duration.

Among all models tested, the **Tuned Random Forest Classifier** delivered the best performance, achieving **99.77% accuracy** and a **ROC-AUC of 0.9892**, making it the ideal model for deployment. This system can help lenders make informed, data-driven decisions and significantly reduce financial risk.