DATA 690

Machine Learning Data Science Project

Michael Good

Professor Jon McKeeby

Predictive Modeling for Credit Card Approval Decisions Using Machine Learning

Dec 4, 2025

Executive Summary

The purpose of this project was to develop a comprehensive credit-risk modeling framework using demographic attributes and historical loan repayment behavior to predict borrower default with accuracy, interpretability, and ethical reliability. The analytical workflow followed a structured approach beginning with extensive data cleaning, exploratory analysis, and preprocessing, all of which ensured that the modeling foundation supported meaningful and trustworthy predictive outcomes. These steps included correcting placeholder values, resolving skewed distributions, normalizing key variables, encoding categorical features, and addressing missing or inconsistent entries that would otherwise compromise model validity.

Early findings from exploratory data analysis revealed significant challenges that shaped the modeling strategy. Income demonstrated a strong right skew, employment duration contained unrealistic placeholder values, and the target variable showed substantial class imbalance between default and non-default cases. Visualizations developed in Unit 5 highlighted these structural concerns and informed several preprocessing decisions, including log transformation, percentile clipping, and the use of SMOTE to rebalance the minority class during training. These enhancements made the dataset more analytically robust and ensured that the predictive models could learn meaningful distinctions between borrower risk profiles.

Predictive modeling in Unit 7 evaluated three machine learning algorithms: Gradient Boosting, XGBoost, and Random Forest. Each model was assessed using a comprehensive suite of performance metrics including ROC-AUC, precision, recall, F1-score, confusion matrices, calibration curves, and precision–recall trade-offs. The results showed clear differentiation in model effectiveness. Gradient Boosting produced the weakest performance with limited discriminatory ability, while XGBoost

demonstrated improved recall and more balanced classification patterns. Random Forest emerged as the strongest overall model, achieving the highest ROC-AUC, the best balance of precision and recall for the minority class, and the most stable probability calibration. These results were reinforced through feature importance analysis and SHAP explanations, which consistently highlighted income, employment duration, and age as the primary drivers of default risk.

The ethical review conducted in Unit 8 confirmed that the modeling approach upheld key standards of fairness, transparency, and responsible data use. Corrections to misleading values, careful handling of class imbalance, and the integration of interpretability tools supported a modeling process aligned with industry and regulatory expectations. Additionally, the project assessed key performance indicators related to model accuracy, risk reduction, and explainability. Random Forest excelled across all KPIs, confirming its suitability as the champion model for credit-risk prediction in this dataset.

Overall, the project successfully delivered a complete and well-validated credit-risk assessment framework. It demonstrates strong methodological execution across data preprocessing, visualization, model development, interpretability, and ethical evaluation. The final Random Forest model provides a reliable, interpretable, and operationally meaningful tool that can support informed and transparent lending decisions.

Table of Contents

Project Scope

Unit 2 Assignment

Problem Description

Credit risk evaluation plays a vital role in maintaining stability within the financial sector. Traditional credit scoring models, such as logistic regression and discriminant analysis, have been widely used to classify applicants as low or high risk (Hand & Henley, 1997). However, these approaches often struggle to capture nonlinear relationships between borrower characteristics and default behavior. As a result, financial institutions increasingly rely on machine learning (ML) models, which can identify complex patterns in large datasets.

Although these advanced models improve predictive accuracy, they introduce challenges related to interpretability and fairness. Financial regulators require institutions to explain credit decisions to consumers, which can be difficult when using opaque algorithms like support vector machines or ensemble methods (Martens et al., 2007). The balance between predictive performance and transparency remains an ongoing concern for banks and credit card issuers.

This project explores predictive modeling for credit card approval decisions using machine learning techniques. It will focus on developing interpretable and accurate classification models to balance predictive performance with transparency and fairness in financial decision-making.

As financial institutions increasingly rely on data-driven automation, ensuring algorithmic fairness and transparency has become critical to prevent bias in lending decisions (Ribeiro et al., 2016). This research aligns with the requirements of the Equal Credit Opportunity Act (ECOA) and Fair Credit Reporting Act (FCRA), which mandate explainable and nondiscriminatory credit evaluation processes.

The research problem that I am analyzing is the development of an interpretable and accurate machine learning model to predict credit card approval outcomes while maintaining fairness and compliance within financial decision-making.

Project Importance

This project is important because credit scoring affects both individual consumers and the broader economy. Credit approval decisions determine access to financial products and influence economic participation. The Federal Trade Commission (2007) highlights that credit scores impact not only loan approval but also insurance rates and other financial services, making accurate and fair scoring systems critical for consumer protection. Accurate and interpretable ML credit models can promote financial inclusion by allowing creditworthy but underrepresented consumers to access affordable credit (Khandani et al., 2010).

From an institutional perspective, improving prediction models can substantially reduce credit losses. Khandani, Kim, and Lo (2010) demonstrated that using ML algorithms to forecast credit card delinquencies increased classification accuracy and reduced potential losses by 6 to 25 percent. This research also showed that better consumer credit modeling helps detect early signs of financial instability, improving systemic risk management. The rise of open banking and alternative data sources further increases the need for interpretable ML methods that regulators and consumers can trust.

I selected this project because it integrates financial analytics and machine learning, aligning with my academic background in finance and data analytics. Futhermore, as a finance and analytics student, I'm passionate about leveraging ML to create fair credit systems. The results of this work can help banks optimize risk assessment while providing fairer access to credit. Ultimately, this project will benefit three groups: (1) financial institutions seeking higher efficiency and lower losses, (2) regulators emphasizing explainable models, and (3) consumers who gain more transparent and equitable credit evaluation.

Review Existing Research in Area

Credit scoring has evolved significantly over the past decades. Hand and Henley (1997) provided one of the earliest comprehensive reviews of statistical methods in credit scoring, emphasizing the role of logistic regression, linear discriminant analysis, and decision trees. As data availability increased, researchers began benchmarking multiple classification algorithms to determine the best predictors of default. Baesens et al. (2003) compared logistic regression, neural networks, and support vector machines (SVM) across eight real-world datasets, finding that both SVM and neural networks achieved strong performance while logistic regression remained competitive due to its interpretability. Lessmann et al. (2015) benchmarked 41 classification algorithms and confirmed that ensemble approaches such as random forests and gradient boosting achieve the highest predictive accuracy in credit scoring, outperforming traditional methods across multiple datasets.

Subsequent studies focused on improving both the accuracy and transparency of ML models. Martens et al. (2007) introduced rule extraction from SVMs to make predictions more interpretable without substantially reducing accuracy. Similarly, Ribeiro et al. (2016) developed the Local Interpretable Model-agnostic Explanations (LIME) framework, which allows users to understand the local reasoning behind individual model predictions. Recent research by Vieira, Barboza, Sobreiro, and Kimura (2019) confirmed that ensemble ML techniques, including random forests and boosting, outperform traditional models when predicting loan defaults, particularly in large datasets. Collectively, this literature demonstrates that while ML models provide better predictive power, explainability remains essential for responsible adoption in financial contexts. However, these models' complexity necessitates post-hoc explanation frameworks such as LIME or SHAP to ensure compliance and user trust.

Review Critical Success Factors (CSFs)

As defined by Bullen and Rockart (1981), critical success factors are the limited number of areas in which satisfactory performance ensures successful outcomes. Ribeiro et al. (2016) emphasized that interpretability is a primary factor for ensuring model trust and adoption in high-stakes environments such as finance. Martens et al. (2007) also noted that transparency is necessary for compliance with credit regulations that require financial institutions to explain adverse decisions to applicants.

Three CSFs for this project include:

- **Data Quality and Preparation –** Reliable model performance depends on accurate, complete, and representative data. Inconsistent or biased data could reduce model validity and fairness.
- **Algorithm Selection and Explainability –** Choosing algorithms that balance predictive accuracy and interpretability, such as logistic regression, decision trees, and explainable AI tools like LIME.
- **Model Evaluation and Validation –** Implementing robust validation techniques (e.g., cross-validation and ROC analysis) to ensure model generalization and reliability.

By addressing these CSFs, the project can produce a model that is both accurate and transparent, thereby meeting both business and regulatory requirements. Additionally, adherence to model risk management standards such as the Federal Reserve's SR 11-7 guidance (Reserve, 2011) is a critical success factor for deploying ML models in production financial environments.

Review Key Performance Indicators (KPIs)

According to Parmenter (2015), key performance indicators (KPIs) are measurable values that reflect how effectively a project meets its objectives. In ML-based credit modeling, these indicators evaluate both model accuracy and its practical benefits.

1. **Model Accuracy and AUC (Area Under Curve):** Evaluates the predictive power of the model in classifying approved versus rejected applicants. Baesens et al. (2003) and Vieira et al. (2019) identified AUC as a standard performance metric in financial modeling.
2. **Reduction in Default Rate:** Measures improvement in identifying high-risk applicants compared to traditional methods, similar to reductions achieved by Khandani et al. (2010).
3. **Transparency and Interpretability:** Tracks whether stakeholders, including analysts and compliance officers, can understand and explain why the model made specific predictions and challenge decisions where necessary (Ribeiro et al., 2016; Martens et al., 2007)

These KPIs ensure that the project's success extends beyond accuracy to include usability and trustworthiness.

Dataset Description

The dataset selected for this project is the Credit Card Approval Prediction dataset from Kaggle (2022). It includes two linked tables: application_record.csv and credit_record.csv. The dataset contains demographic and financial variables for credit card applicants, such as income, education level, marital status, housing type, and employment duration. The credit_record.csv table includes monthly payment statuses that can be merged with application data using a common client ID. The dataset will be preprocessed using label encoding and imputation techniques to handle missing values and categorical variables, ensuring data consistency for supervised learning.

This dataset enables the construction of a binary classification model that predicts whether an applicant will be approved or denied for a credit card based on historical behavioral and financial patterns. It contains over 400,000 records, making it suitable for training complex ML models. The presence of missing values, categorical features, and class imbalance reflects real-world data challenges in credit risk modeling.

The dataset holds a perfect Kaggle quality score (10/10), indicating full completeness, credibility, and compatibility. This means it includes verified provenance, consistent update frequency, detailed feature documentation, and standardized file formatting, making it highly reliable for academic and applied data analytics research.

Data Analytics Tools

The primary analytical tools for this project include Python, Scikit-learn, Pandas, and Matplotlib for data preprocessing, model training, and visualization. These tools support algorithms such as logistic regression, decision trees, random forests, and SVMs. LIME will be used to interpret individual predictions and enhance transparency (Ribeiro et al., 2016). Model performance will be validated using stratified k-fold cross-validation to ensure generalizability and reduce overfitting.

For visualization and reporting, Tableau or Power BI may be used to create interactive dashboards that summarize model performance, feature importance, and key insights. These tools help communicate results effectively to both technical and non-technical audiences, ensuring clarity in the decision-making process.

All analyses will be conducted in Google Colab, utilizing its cloud-based Python environment with integrated Jupyter notebooks for code execution, visualization, and reproducibility. This platform facilitates seamless collaboration, version control via Google Drive, and access to GPU-accelerated computation for model training.

Project Milestones – Lessons Learned

Describe the major milestones and the lessons you learned for each assignment.

Unit 2 Assignment Project Scope	**Unit 1 – Project Selection:** Defined the project's direction by selecting the Credit Card Approval Prediction dataset from Kaggle, developing a measurable research problem, and outlining the project's goals. Established the connection between predictive modeling, financial decision-making, and model interpretability. Created the initial framework that links machine learning techniques with the practical challenges of credit approval systems. **Unit 2 – Research Foundation:** Conducted an extensive literature review of peer-reviewed research and industry publications on credit scoring, machine learning methods, and explainable AI. Analyzed findings from scholars such as Hand and Henley (1997), Martens et al. (2007), Ribeiro et al. (2016), and Vieira et al. (2019), focusing on how each contributes to the evolution of predictive modeling and model transparency. Evaluated the strengths and limitations of logistic regression, support vector machines, and ensemble methods for credit risk assessment.	I learned the importance of defining clear and measurable problem boundaries to maintain focus throughout the project. This stage emphasized how critical it is to align the dataset and objectives with a real-world application that has both technical and ethical implications. I also realized that an effective capstone project requires more than technical execution, but it must demonstrate relevance to business objectives, data governance, and social responsibility. By articulating the problem early, I set a foundation that ensures future steps, such as model development and evaluation, will be purposeful and cohesive. I learned that the most successful analytics projects are not solely defined by high model accuracy but by their ability to provide transparent, fair, and actionable insights. Through this research, I gained a deeper understanding of how interpretability techniques such as LIME can improve trust in machine learning models, particularly in regulated financial settings. Reviewing these studies reinforced that balancing predictive power with ethical considerations is essential for sustainable analytics. This stage also improved my ability to critically analyze academic literature and translate theoretical findings into practical project design choices.
Unit 4 Assignment Data Selection and Analysis	**Unit 4 – Selection and Analysis** In Unit 4, I conducted a detailed data validation and exploratory analysis of the two datasets ***application_record.csv*** and ***credit_record.csv***. Using Python and pandas, I verified dataset completeness, identified missing values, and calculated summary statistics for key numerical variables such as income, age, and employment duration. I transformed raw values into interpretable metrics, such as converting DAYS_BIRTH and DAYS_EMPLOYED into positive years and replaced placeholder values (365,243) representing unemployed individuals. I also examined categorical distributions and the frequency of credit statuses to understand repayment patterns across 45,985 unique clients. These steps ensured that both datasets	Through this stage, I learned the importance of performing a thorough data audit before developing predictive models. Even a high-quality dataset, like this one, can contain encoded or misrepresented values that distort results if left unaddressed. By applying Python-based checks for completeness and descriptive statistics, I strengthened my ability to detect and correct inconsistencies systematically. I also gained deeper insight into how demographic and behavioral attributes interact in credit risk modeling. This process reinforced the value of preprocessing, feature understanding, and documentation to maintain transparency and reproducibility in machine learning workflows.

	were accurate, consistent, and ready for modeling in subsequent units.	
Unit 5 Assignment Presentation	**Unit 5 – Presentation and Peer Review** In Unit 5, I created and delivered a narrated PowerPoint presentation summarizing the current status of my predictive modeling capstone. This included developing slides covering the research problem, project scope, the importance of the project, dataset descriptions, and lessons learned from preprocessing and visualization work. I prepared detailed narration for each slide and ensured that the presentation aligned with course expectations by focusing on data cleaning challenges, exploratory findings, and how these steps support future modeling work.	Through this assignment, I learned how to communicate technical progress clearly and effectively to an audience. Preparing the presentation reinforced the importance of structuring a data science project explanation so that each section builds logically toward the modeling phase. It also helped me articulate how issues such as skewed income, placeholder values, and class imbalance directly affect model reliability and fairness. Overall, this assignment strengthened my ability to translate analytical tasks into clear, interpretable insights suitable for peer review and professional presentation.
Unit 5 Assignment Visualization	**Unit 5: Assignment Visualization** In Unit 5, I created three exploratory visualizations to assess key variables in the credit scoring dataset and evaluate how preprocessing choices influence statistical interpretation. I generated a dual-panel histogram for income that compared log transformation against 99th percentile clipping to address strong right skew. I also produced a KDE-enhanced histogram to examine the age distribution of applicants and a labeled bar chart summarizing the frequency of credit status categories across 45,985 clients. These visualizations required preparing the underlying data by removing placeholder values, scaling skewed features, and recoding categorical attributes. Together, the three figures provided a structured exploratory view of income patterns, demographic characteristics, and repayment behavior.	Through this assignment, I learned how essential visualization is for diagnosing distributional issues such as skew, outliers, and class imbalance before developing predictive models. The contrast between raw and transformed income data showed how preprocessing decisions can significantly alter the clarity and interpretability of a variable. Examining the age distribution improved my understanding of demographic patterns in the dataset, while the credit status chart highlighted the severity of class imbalance and its implications for model evaluation. Overall, this unit strengthened my ability to use visual analysis to guide preprocessing strategy, feature engineering, and the preparation required to build reliable machine learning models.
Unit 7 Assignment Predictive Modeling	**Unit 7 Assignment: Predictive Modeling** In Unit 7, I developed and evaluated three supervised machine learning models to predict credit default using a highly imbalanced dataset. I prepared the modeling pipeline by standardizing features, splitting the data into training and test sets, and applying SMOTE to address minority-class underrepresentation. I	Through this assignment, I learned how significantly class imbalance affects predictive modeling performance and why imbalance-sensitive methods such as SMOTE and ensemble algorithms are essential in credit-risk applications. Evaluating Gradient Boosting, XGBoost, and Random Forest side-by-side demonstrated how different tree-based algorithms respond to the same data and why some

	then trained Gradient Boosting, XGBoost, and Random Forest models, each selected for their ability to capture non-linear patterns relevant to credit risk. For every model, I generated a full suite of evaluation outputs including confusion matrices, ROC curves, precision–recall curves, calibration plots, and classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. I also performed interpretability analysis using feature importance measures and SHAP values for XGBoost and Random Forest to understand which borrower characteristics most strongly influenced model predictions. After evaluating performance across all models, I compared them systematically to identify the champion model. Based on discriminative ability, probability calibration, and minority-class performance, the Random Forest model emerged as the strongest overall and was formally selected as the project's champion.	models generalize better under noisy, asymmetric class distributions. I also learned how to interpret model behavior using SHAP, which improved my understanding of how features such as income, age, and years employed contribute to the predicted probability of default. Additionally, this assignment strengthened my ability to use multiple performance metrics rather than relying solely on accuracy. Metrics like ROC-AUC, precision, recall, and calibration curves showed how a model can appear strong on overall accuracy while still failing to identify high-risk borrowers. This helped reinforce the importance of using domain-appropriate evaluation criteria when working with financial risk models. Ultimately, the assignment improved my ability to build, interpret, and compare predictive models in a way that balances both performance and practical relevance.
Unit 8 Assignment Final Report Findings	**Unit 8 Assignment: Final Report** For Unit 8, I compiled all prior project components into a complete final report, integrating findings from data cleaning, visualization, preprocessing, and model evaluation. I summarized the performance of all three predictive models, justified the Random Forest as the champion model, and incorporated tables and interpretability results to support the conclusions. I also wrote the required sections on ethical considerations, project success, KPIs, lessons learned, and future analysis to finalize the end-to-end modeling workflow.	Completing this unit showed me how to bring together all technical and analytical steps into a cohesive narrative that clearly communicates project results. I learned the importance of connecting model outputs to real-world implications, including fairness, transparency, and responsible data use in credit modeling. This unit reinforced the value of structured reporting and demonstrated how strong documentation and ethical awareness are essential parts of deploying machine learning solutions.

Data Selection

Unit 4 Assignment

Data Summary

The dataset selected for this capstone project is titled Credit Card Approval Prediction and is available publicly on Kaggle (Kaggle, 2022). It consists of two files being *application_record.csv* and *credit_record.csv* that are joined using a shared client identifier labeled “ID.” The first file captures demographic and financial attributes of credit card applicants, while the second file records their monthly credit repayment history. Together, these two datasets form a robust foundation for building and evaluating machine learning models to predict credit approval outcomes.

This dataset is highly valuable because it provides an opportunity to study the demographic, socioeconomic, and behavioral factors that influence creditworthiness. Financial institutions rely heavily on predictive models to evaluate risk, but ensuring these models are both accurate and transparent remains a challenge. By analyzing this dataset, the project will explore how explainable machine learning methods can support fairer and more data-driven credit decisions. The information contained within the dataset closely reflects real-world data used in credit evaluation, making it an ideal foundation for supervised learning and model interpretability research.

Both files were collected and anonymized by the contributor “rikdifos” on Kaggle, who compiled financial records from credit applicants. The application_record.csv file contains 438,557 rows and 18 columns, while the credit_record.csv file includes 1,048,575 rows and 3 columns. These files cover several years of applicant data, with the credit record using monthly time-based entries to track borrower performance. To ensure proper model validation, the dataset will be split into 70 percent for training and 30 percent for testing to assess generalization performance and minimize overfitting.

A data completeness analysis was conducted in Google Colab using Python's *pandas* library. The evaluation revealed that application_record.csv is 98.3 percent complete, with missing values only in the OCCUPATION_TYPE field, while credit_record.csv is 100 percent complete. This confirms that the dataset is of high quality and suitable for predictive modeling, with only minor imputation or categorical handling required before preprocessing.

Dataset citation (APA 7th edition):

Rikdifos. (2022). *Credit Card Approval Prediction* [Data Set]. Www.kaggle.com. https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

Data Definition/Data Profile

The dataset is divided into two related files. The first file, application_record.csv, contains information about each applicant's demographics, income, family composition, and contact availability. The second file, credit_record.csv, includes monthly-level data reflecting each applicant's payment history and account status. The following tables describe each field, its data type, and potential data quality concerns identified during preliminary profiling.

Table 1. Application Record Variables

Field Name	Definition / Description	Data Type	Potential Outliers	Frequency of Nulls	Potential Quality Issues
ID	Unique client identifier used to link both datasets	Integer	None	0%	None
CODE_GENDER	Gender of applicant	Categorical	Imbalanced distribution	0%	Minor missing values
FLAG_OWN_CAR	Indicates car ownership	Binary (Y/N)	None	0%	None
FLAG_OWN_REALTY	Indicates property ownership	Binary (Y/N)	None	0%	None
CNT_CHILDREN	Number of dependent children	Integer	Large families (>6)	0%	Some extreme values
AMT_INCOME_TOTAL	Total annual income	Float	Very high income values	0%	Possible income outliers
NAME_INCOME_TYPE	Income source (e.g., working, pensioner, student)	Categorical	None	0%	Possible mislabeling
NAME_EDUCATION_TYPE	Highest education level	Categorical	None	0%	None

NAME_FAMILY_STATUS	Marital status	Categorical	None	0%	None
NAME_HOUSING_TYPE	Type of housing (own, rent, etc.)	Categorical	None	0%	Inconsistent labeling
DAYS_BIRTH	Age of applicant in days (negative = days since birth)	Integer	Older ages	0%	Must convert to years
DAYS_EMPLOYED	Employment duration in days (negative = employed)	Integer	Values >0 = unemployed	0%	Sign interpretation
FLAG_MOBIL	Indicates mobile phone ownership	Binary (0/1)	None	0%	None
FLAG_WORK_PHONE	Indicates work phone availability	Binary (0/1)	None	0%	None
FLAG_PHONE	Indicates home phone ownership	Binary (0/1)	None	0%	None
FLAG_EMAIL	Indicates email address availability	Binary (0/1)	None	0%	None
OCCUPATION_TYPE	Applicant's occupation type	Categorical	Rare categories	30.6%	Missing or anonymized data
CNT_FAM_MEMBERS	Family size	Integer	Large households (>7)	0%	Rounding inconsistencies

Table 2. Credit Record Variables

Field Name	Definition / Description	Data Type	Potential Outliers	Frequency of Nulls	Potential Quality Issues
ID	Unique client identifier	Integer	None	0%	None
MONTHS_BALANCE	Record month (0 = current, -1 = previous, etc.)	Integer	None	0%	None
STATUS	Monthly payment status (0–5, C, X)	Categorical	Long delinquency streaks	0%	May require numeric recoding

Descriptive Statistics

Descriptive statistics are used to summarize, organize, and interpret the characteristics of a dataset through measures such as mean, median, and standard deviation. According to Kaur, Stoltzfus, and Yellapu (2018), descriptive statistics are essential for providing a clear understanding of data patterns and variability before performing inferential or predictive analysis. These measures of central tendency and dispersion help identify potential outliers, reveal the shape of data distributions, and guide preprocessing decisions. Amrhein, Trafimow, and Greenland (2019) further emphasize that descriptive statistics are foundational representations of observed data rather than inferential generalizations. They argue that descriptive analysis clarifies the structure and relationships within the data, ensuring interpretations are based on empirical evidence rather than assumptions.

Using Python's *pandas* library, descriptive statistics were computed for all numeric variables in application_record.csv. The dataset contains 438,557 observations, and the summary statistics reveal strong diversity among applicants. The average annual income (AMT_INCOME_TOTAL) is approximately 187,524 units with a standard deviation of 110,087, indicating wide variation in earnings across applicants. The median number of children (CNT_CHILDREN) is 0, showing that most applicants do not have dependents, while the average family size (CNT_FAM_MEMBERS) is 2.19, reflecting smaller households. The average applicant age (DAYS_BIRTH) is about 43.8 years, derived by converting negative day counts into positive years. Initially, the average employment duration (DAYS_EMPLOYED) appeared unrealistically high at 165.9 years due to placeholder values (365,243) representing unemployed applicants. These values were corrected by replacing them with nulls and converting valid negative day counts into positive years, resulting in a realistic average employment duration of 7.2 years. This correction ensures that employment history is accurately represented before model training.

The dataset's skewness and kurtosis measures reveal several insights. For example, AMT_INCOME_TOTAL exhibits strong right skew (8.83) and high kurtosis (324.55), suggesting that a few high-income outliers disproportionately affect the distribution. In contrast, DAYS_BIRTH is approximately symmetric (skew = −0.16), indicating a balanced age spread among applicants. Binary indicators such as FLAG_MOBIL, FLAG_WORK_PHONE, and FLAG_EMAIL display limited variance, which may lower their influence in predictive modeling. These findings highlight the need for normalization and potential feature scaling to enhance model performance and interpretability.

The credit_record.csv dataset complements the applicant demographic data by providing longitudinal payment histories for each client. It contains 1,048,575 records across 45,985 unique clients, with three variables: ID, MONTHS_BALANCE, and STATUS. The STATUS field captures credit repayment behavior over time, using categorical values where "0" indicates 1–29 days past due, "1"–"5" reflect increasing delinquency severity, "C" represents accounts paid off that month, and "X" denotes months without active loans. Frequency analysis revealed that most records are "C" (442,031) or "0" (383,120), showing that the majority of applicants maintain timely or closed accounts, while higher delinquency statuses ("3"– "5") occur infrequently. The dataset demonstrates complete records with no missing values, confirming its reliability for constructing the target variable that distinguishes between "good" and "bad" clients. Because the variables are categorical and time-dependent, traditional descriptive measures such as mean or standard deviation are not applicable. Instead, frequency distributions and temporal transition patterns will be used to evaluate repayment consistency and risk trends.

Overall, these descriptive insights demonstrate that the dataset captures a realistic range of demographic and financial characteristics. The diversity in applicant profiles strengthens the potential generalizability of predictive models, while the preprocessing corrections to employment data ensure data integrity. Together, these findings establish a strong analytical foundation for developing transparent and data-driven credit scoring models that balance predictive accuracy with interpretability.

Table 3. Descriptive Statistics for Numeric Fields

Feature name	count	mean	std	min	25%	50%	75%	max	skew	kurtosis
ID	438557.0	6022176.27	571637.02	5008804.0	5609375.0	6047745.0	6456971.0	7999952.0	0.21	-0.09
CNT_CHILDREN	438557.0	0.43	0.72	0.0	0.0	0.0	1.0	19.0	1.81	5.08
AMT_INCOME_TOTAL	438557.0	187524.29	110086.85	26100.0	121500.0	160780.5	225000.0	6750000.0	8.83	324.55
DAYS_BIRTH	438557.0	-15997.90	4185.03	-25201.0	-19483.0	-15630.0	-12514.0	-7489.0	-0.16	-1.05

DAYS_EMPLOYED	438557.0	60563.68	138767.80	-17531.0	-3103.0	-1467.0	-371.0	365243.0	1.74	1.03
FLAG_MOBIL	438557.0	1.00	0.00	1.0	1.0	1.0	1.0	1.0	0.00	0.00
FLAG_WORK_PHONE	438557.0	0.21	0.40	0.0	0.0	0.0	0.0	1.0	1.45	0.11
FLAG_PHONE	438557.0	0.29	0.45	0.0	0.0	0.0	1.0	1.0	0.94	-1.12
FLAG_EMAIL	438557.0	0.11	0.31	0.0	0.0	0.0	0.0	1.0	2.52	4.36
CNT_FAM_MEMBERS	438557.0	2.19	0.90	1.0	2.0	2.0	3.0	20.0	0.92	1.90

Note. This table summarizes the distribution characteristics of all numeric variables in the application_record.csv dataset, including count, mean, standard deviation, quartiles, and measures of skewness and kurtosis.

Data Visualizations

Unit 5 Assignment

Data Visualization Definitions

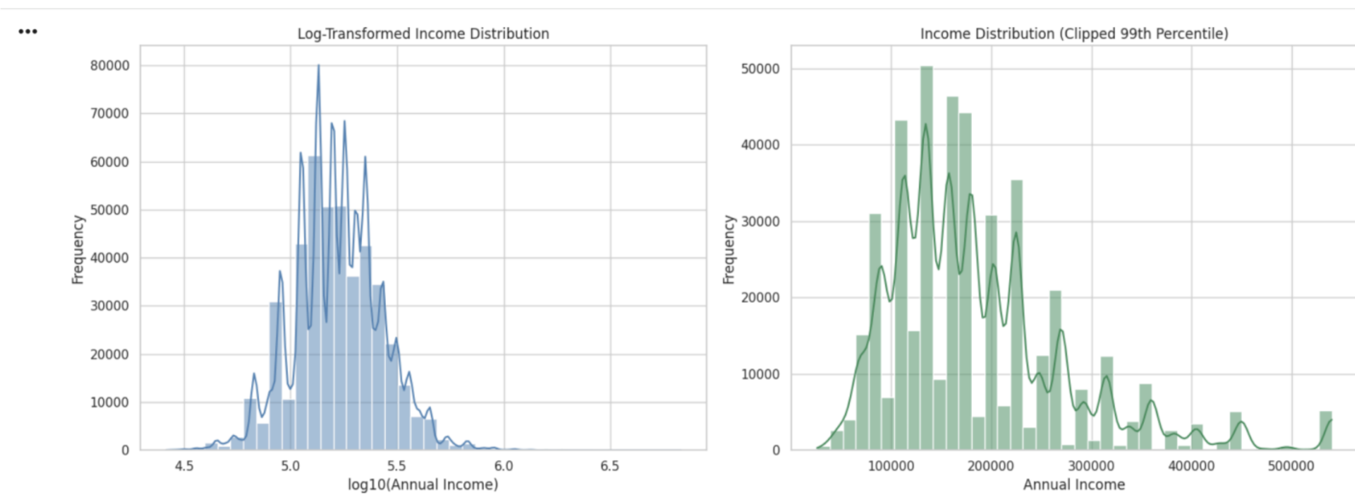
This project uses three core visualization techniques to explore patterns in applicant demographics and credit behavior: histograms, kernel density estimation, and bar charts. A histogram is a foundational tool in exploratory data analysis because it groups continuous variables into bins and visually displays the frequency distribution of the data. Tukey (1977) emphasized that histograms allow analysts to observe unexpected structure in the data, such as skewness or clustering, that may not be visible through descriptive statistics alone. In this project, histograms are used to examine the distribution of key financial variables such as annual income and age, allowing for the detection of extreme values, skewed distributions, and unrealistic outliers that may affect later modeling stages.

Kernel density estimation (KDE) is used alongside histograms to generate a smooth probability curve that represents the underlying distribution without relying on rigid bin boundaries. KDE helps reveal subtle patterns in the data that can be obscured by uneven or coarse binning choices. Gelman (2004) notes that density estimation supports deeper understanding of distributional shape by reducing noise and uncertainty, especially in complex datasets where variability is high. Mukhiya and Ahmed (2020) emphasize that visual exploration in Python, particularly with histograms and density plots, is essential for identifying anomalies, understanding distributional behavior, and guiding preprocessing decisions in machine learning workflows. KDE is applied here to complement histogram analysis by smoothing the income and age distributions, enabling clearer interpretation of central tendencies, long tails, and the influence of extreme values such as top one percent earners.

Bar charts are the final visualization technique incorporated in this project. They are used to summarize categorical credit outcomes, such as monthly repayment status codes. Bar charts enable straightforward comparison of category frequencies and highlight class imbalance, which is a critical issue in credit scoring applications. Yim, Chung, and Yu (2018) describe bar charts as essential for summarizing categorical variables because they provide a clear visual representation of magnitude across discrete groups. In the context of this dataset, bar charts allow for easy identification of the dominant repayment categories, the prevalence of delinquency indicators, and the overall distribution of credit behavior across the sample.

Together, these visualization techniques support a comprehensive exploratory analysis. Histograms and KDE curves reveal the structure of continuous financial attributes, while bar charts provide insight into categorical repayment patterns. Using these complementary methods ensures a thorough understanding of the dataset before moving forward to predictive modeling and feature engineering.

Data Visualization 1



Description of the Visualization

Figure 1 presents a side-by-side comparison of applicant income distributions using two preprocessing approaches: a log-transformed histogram with KDE (left) and a percentile-clipped histogram with KDE (right). The left panel displays the distribution of AMT_INCOME_TOTAL after applying a base-10 logarithmic transformation, which compresses the high-income range and allows the rounded mass of the distribution to become more visible. The right panel shows the same income variable in raw currency terms but clipped at the 99th percentile to limit the influence of extreme outliers. Both histograms include overlaid kernel density curves to provide a smoothed representation of the distribution's shape.

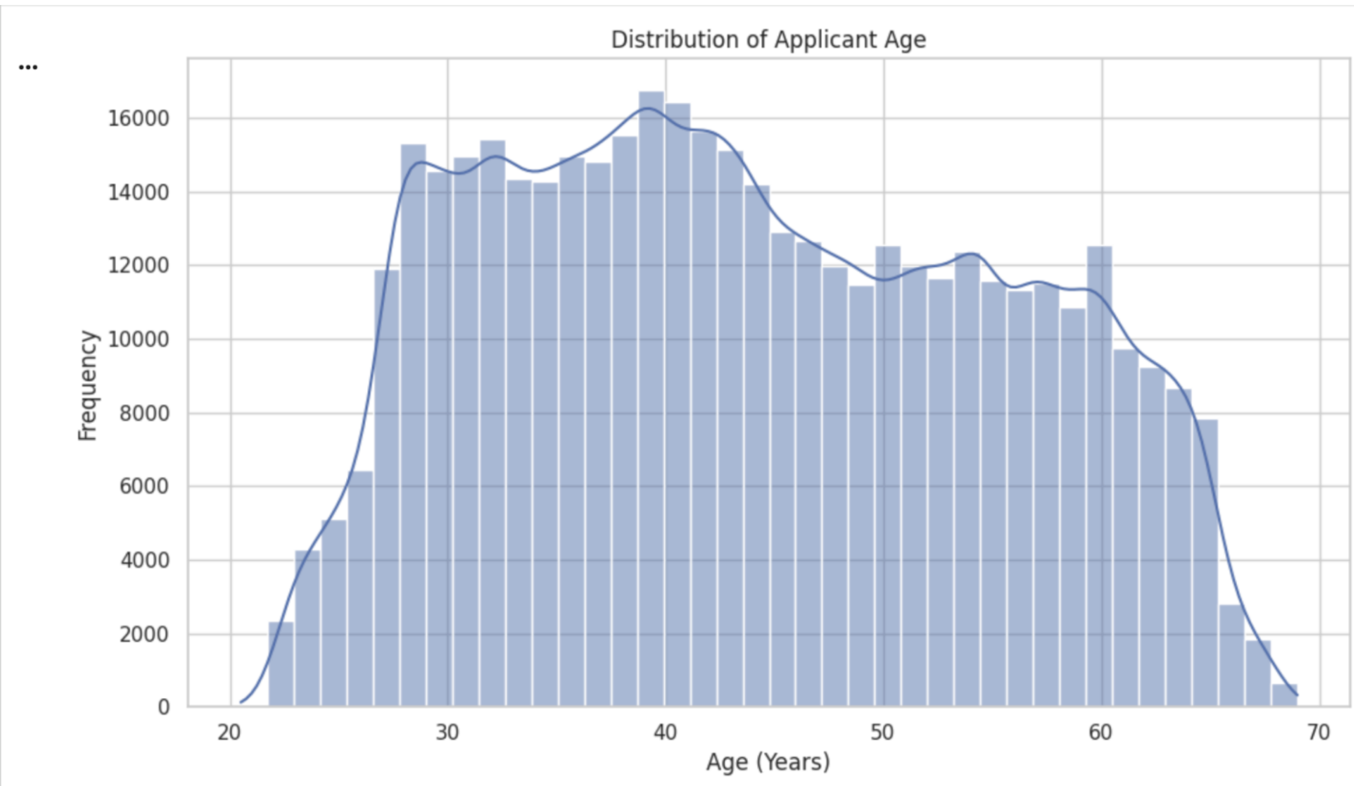
Without preprocessing, the raw income distribution becomes nearly unreadable. The unprocessed histogram collapses the majority of applicants into a narrow region near zero on the x-axis due to a very long upper tail that stretches into the millions. This extreme compression conceals central tendencies, density structure, and meaningful variance in the data. By applying log transformation and percentile clipping, the visualization corrects this distortion and reveals the true underlying patterns in the dataset.

Insights From the Visualization

The first key insight from this visualization is that the income variable exhibits strong right skew when displayed in its raw scale, even after clipping extreme observations. The histogram on the right shows a heavy concentration of applicants earning between roughly 50,000 and 250,000 units of currency, while the KDE curve gradually tapers across the higher-income range. This long tail suggests that a relatively small portion of applicants earn substantially more than the majority—a pattern that is characteristic of financial and socioeconomic variables. The use of clipping helps mitigate the influence of extreme values, but the skew remains visible, indicating that transformations or scaling techniques may be necessary for downstream modeling.

The second insight emerges through the log-transformed panel. After applying a log transformation, the distribution becomes far more symmetric, with a clear unimodal peak and noticeably reduced spread. This demonstrates that the log transformation stabilizes variance, compresses extreme values, and produces a shape that is more suitable for statistical modeling and machine learning algorithms that assume roughly normal input features. The KDE curve in the log scale provides a smooth, bell-shaped structure, indicating that log transforming income successfully enhances interpretability. Together, the two panels highlight why preprocessing choices materially change the representation of skewed financial variables and support decisions about which transformation is best aligned with analytical goals.

Data Visualization 2



Description of the Visualization

Visualization 2 presents a histogram with an overlaid kernel density estimate (KDE) illustrating the distribution of applicant ages in the dataset. The x-axis represents applicant age in years, while the y-axis shows the frequency of applicants within each age bin. The histogram bars provide a discrete count of applicants by age group, and the KDE curve adds a smoothed approximation of the overall age density.

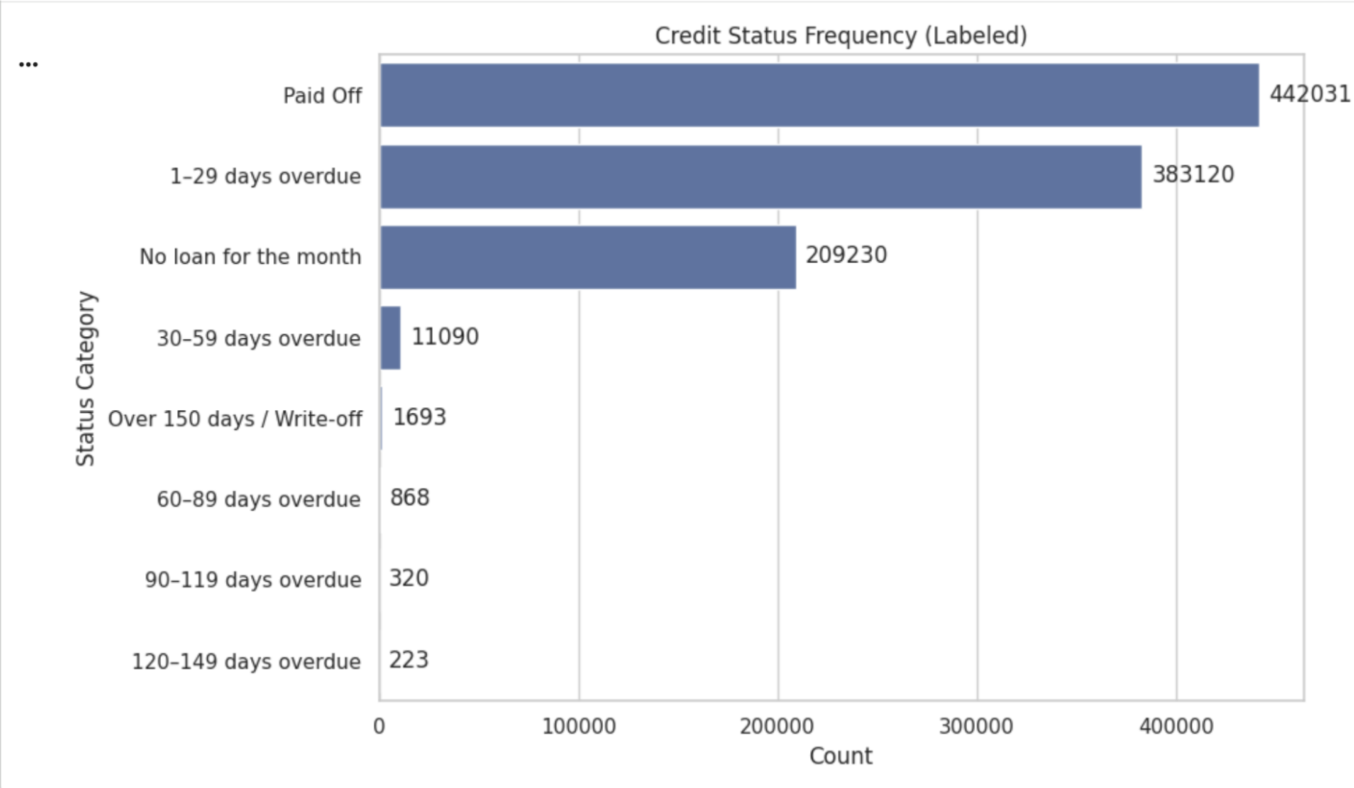
Together, these elements give a clear sense of how applicants are spread across the age spectrum, highlighting the shape, concentration, and variability of the demographic distribution.

Insights From the Visualization

A primary insight from this visualization is that the applicant population is not uniformly distributed across age groups; instead, it clusters heavily between the late 20s and mid-40s. The KDE curve peaks prominently in the early-to-mid-30s, indicating that this age group represents the highest concentration of applicants. This type of distribution is common in consumer finance datasets, where individuals in their working-age prime tend to engage most actively in credit-seeking and loan-related activities. The right side of the distribution gradually tapers from the mid-40s through the early-60s, showing steady participation but decreasing frequency as age increases.

A second insight is that the distribution exhibits a mild left tail and a more extended right tail, reflecting lower participation among younger and older applicants. Few individuals under age 25 appear in the dataset, which may indicate either lower eligibility or limited credit activity in that demographic. On the other end, applicants above 60 show a visible decline in frequency, which is consistent with typical patterns of reduced borrowing behavior approaching retirement age. The smoothness of the KDE curve also suggests that the age variable does not suffer from extreme outliers or irregularities, making it a clean and reliable feature for downstream modeling. Overall, the visualization reveals a well-structured, moderately skewed age distribution dominated by applicants in their peak earning and borrowing years.

Data Visualization 3



Description of the Visualization

Visualization 3 presents a horizontal bar chart that displays the frequency of applicants across different labeled credit status categories. Each bar shows the total number of observations that fall into a specific repayment or delinquency group, including Paid Off, No loan for the month, and several overdue ranges. Numeric labels are placed at the end of each bar to show the exact count of applicants in each category. The horizontal layout keeps the category names readable and makes it easy to compare the length of the bars.

Insights From the Visualization

A key insight from this visualization is that the majority of applicants fall into positive or minimally delinquent credit categories. The Paid Off group is the most common with 442,031 applicants, followed by the 1 to 29 days overdue group with 383,120 applicants, and the No loan for the month group with 209,230 applicants. These three categories make up most of the dataset, which suggests that most individuals either repay their loans on time, experience only short delays, or do not have an active loan during the reported month. This concentration also shows that the dataset is imbalanced toward lower risk credit behavior, which can influence how predictive models interpret repayment patterns.

Another important insight is the sharp decline in frequency as the severity of delinquency increases. The 30 to 59 days overdue category drops to 11,090 applicants, and the deeper delinquency groups decrease even further, with some categories containing only a few hundred observations. The most severe category, Over 150 days or Write-off, includes only 1,693 records relative to the overall dataset. This pattern reflects typical financial behavior where only a small portion of borrowers reach serious delinquency or default. Although this is positive from a lending perspective, it introduces challenges for credit risk modeling because the most critical classes are rare. These smaller groups may require techniques such as oversampling or alternative evaluation metrics to ensure accurate prediction. Overall, the visualization highlights both the distribution of repayment behavior and the analytical considerations that come with studying rare but important credit outcomes.

Summary

The three visualizations collectively provide a clearer understanding of applicant characteristics and repayment behavior. The income distributions illustrate that raw financial variables are highly skewed and dominated by outliers, making preprocessing steps such as log transformation and percentile clipping essential for interpretability. These transformations reveal the true central pattern of applicant incomes and highlight the importance of scaling and normalization during model preparation. The age distribution adds another demographic perspective, showing a smooth and unimodal pattern concentrated across mid-career age groups, which indicates that age is a stable and informative feature for downstream modeling.

The credit status bar chart further reveals significant class imbalance, with most applicants falling into either "Paid Off" or "1 to 29 days overdue," while severe delinquency is rare. This imbalance has direct implications for predictive modeling because unbalanced outcomes often require adjusted evaluation metrics or resampling techniques to ensure accurate risk identification. Together, the three visualizations establish a foundation for understanding the dataset's structure and statistical challenges. These findings align with the literature, which emphasizes that preprocessing and visualization are essential for identifying outliers, understanding variable behavior, and informing analytical strategy (Mukhiya and Ahmed, 2020).

Predictive Models

Unit 7 Assignment

Predictive Modeling Definitions

This project incorporates several supervised machine learning techniques to evaluate and predict credit default behavior using a labeled dataset of consumer credit histories and demographic features. Although four different models were explored during experimentation, the formal predictive models selected for full analysis and comparison are Random Forest, Gradient Boosting, and XGBoost. Logistic Regression was included early as a baseline classifier due to its simplicity and interpretability, and its results provided a useful benchmark for assessing the value added by more complex ensemble methods. Logistic Regression is a probabilistic linear classifier that models the relationship between input features and the probability of a binary outcome. It generates interpretable coefficient estimates and serves as a baseline model in many credit risk studies because it is transparent and aligns well with regulatory expectations for explainability (Hosmer et al., 2013). Although logistic models are limited in their ability to capture nonlinear effects, they provide a useful reference point for comparing the performance of more complex algorithms. As four models were initially trained (Logistic Regression, Random Forest, Gradient Boosting, and XGBoost), Logistic Regression showed substantially lower discriminatory performance (ROC–AUC = 0.5189) which is the basis of its exclusion from the three models selected.

The primary models selected for predictive evaluation were three tree-based ensemble methods that have shown strong performance in credit scoring and financial risk prediction. The first is Random Forest, an algorithm that builds a large number of decision trees and aggregates their outputs to reduce overfitting and improve predictive stability. Breiman (2001) describes Random Forests as a method that increases accuracy by injecting randomness into both feature selection and sampling, which allows the model to capture different patterns across subsets of the data. The second model is Gradient Boosting, which builds trees sequentially where each new tree corrects the errors of the previous ones. Gradient

Boosting has proven especially effective for complex classification problems because it optimizes the model through iterative loss reduction. The third and most advanced model used in this project is XGBoost, a scalable and highly regularized boosting technique designed for both speed and performance. Chen and Guestrin (2016) highlight XGBoost as a method capable of handling sparse data, nonlinear patterns, and large-scale tasks while maintaining strong generalization. The relevance of boosted tree models in financial risk classification is further supported by Carmona et al. (2019), who demonstrated that boosting substantially improves predictive accuracy when modeling default risk in the United States banking sector. Together, these three ensemble algorithms provide a comprehensive and rigorous framework for evaluating credit default patterns within this dataset.

Predictive Model 1

Gradient Boosting was the first supervised learning model evaluated on the credit-default dataset. Gradient boosting builds an additive sequence of weak decision trees, where each tree corrects the residual errors left by the previous ones. This approach allows the model to capture nonlinear patterns in borrower characteristics and credit behavior. Research in financial-risk modeling, such as Carmona et al. (2019), shows that boosting methods are effective for identifying subtle risk signals but can be sensitive to class imbalance, especially when defaults represent a small portion of observations. To address this imbalance, the model in this project was trained on SMOTE-balanced data and evaluated on an untouched test set, ensuring that performance metrics reflected real-world default rates rather than artificially balanced conditions.

The Gradient Boosting Classifier produced mixed results. Its accuracy reached 0.689, but accuracy is misleading given the significant class imbalance. More informative metrics showed limited predictive strength: AUC = 0.559, precision for defaulters was 0.14, recall was 0.32, and the F1-score was 0.20. The ROC curve shows only marginal improvement over random chance, and the precision–recall curve reveals that precision drops quickly as recall increases, which is an expected outcome when the model struggles to distinguish minority class patterns. The confusion matrix highlights the same issue: while the model correctly identified 275 defaults, it misclassified 1,684 non-defaults as defaults (false positives), indicating weak discriminatory ability. The calibration curve shows irregular probability estimates, signaling poor probability reliability for risk scoring.

A deeper examination of the model's error patterns shows that Gradient Boosting tended to overfit to the SMOTE-balanced training data while failing to generalize effectively to the true class distribution. The large number of false positives suggests the model learned broad, overly aggressive decision boundaries that did not translate into dependable real-world predictions. Because the default class is rare, the model appeared to overcompensate by predicting default too frequently, which reduced precision and created unstable risk estimates. This behavior reflects a common limitation of boosting algorithms in imbalanced classification tasks, where their sequential corrections can amplify noise instead of uncovering meaningful structure.

Overall, Gradient Boosting served as a useful baseline among tree-based boosting approaches, but its performance indicates difficulty capturing the complex decision boundaries needed for accurate credit-default detection. While it highlighted general data trends and helped set expectations for boosting performance, its limited discriminative power and weak probability calibration made it unsuitable as a final predictive model. Both XGBoost and Random Forest ultimately outperformed it across all major evaluation metrics, particularly AUC, recall, precision for the minority class, and calibration reliability.

Figure 1. ROC Curve – Gradient Boosting

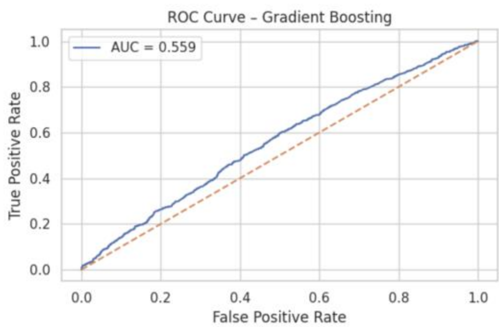
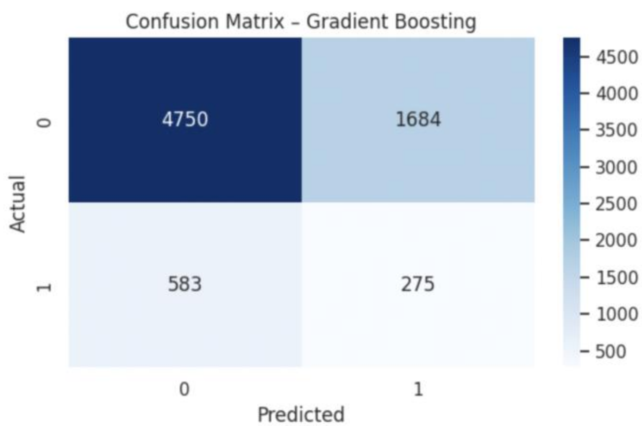


Figure 2. Confusion Matrix – Gradient Boosting




```
              precision    recall  f1-score   support

           0       0.89      0.74      0.81      6434
           1       0.14      0.32      0.20       858

    accuracy                           0.69      7292
   macro avg       0.52      0.53      0.50      7292
weighted avg       0.80      0.69      0.74      7292
```

Figure 3. Train / Test – Gradient Boosting

```
... ===== Training Gradient Boosting =====

    --------------------------------------
    GradientBoostingClassifier
    --------------------------------------
    Accuracy: 0.6891
    Precision: 0.1404
    Recall: 0.3205
    F1 Score: 0.1952
    ROC-AUC: 0.5592

    Confusion Matrix:
    [[4750 1684]
     [ 583  275]]

    Classification Report:
                  precision    recall  f1-score   support

               0       0.89      0.74      0.81      6434
               1       0.14      0.32      0.20       858

        accuracy                           0.69      7292
       macro avg       0.52      0.53      0.50      7292
    weighted avg       0.80      0.69      0.74      7292
```

Figure 4. Calibration Curve – Gradient Boosting

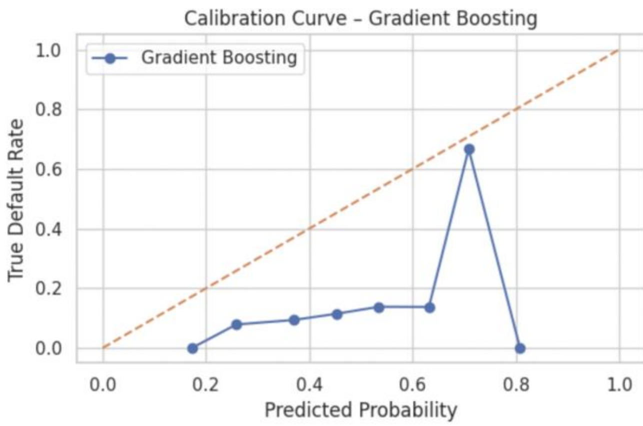
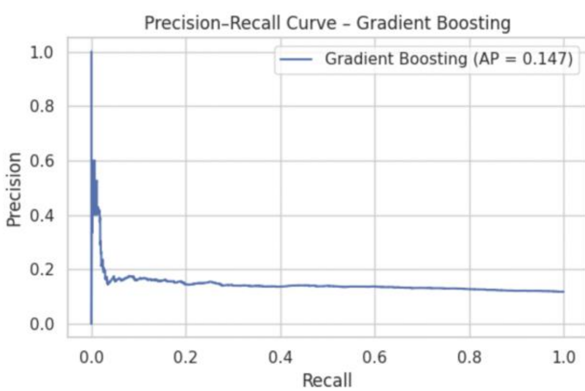


Figure 5. Precision–Recall Curve – Gradient Boosting



Predictive Model 2

The second predictive model applied in this study was Extreme Gradient Boosting (XGBoost), a high-performance tree-boosting algorithm known for its scalability, parallelization, and strong predictive accuracy in structured tabular datasets. XGBoost builds sequential decision trees that correct the residual errors of previous trees, reinforcing high-signal patterns while minimizing overfitting through regularization and shrinkage. Chen and Guestrin (2016) demonstrate that XGBoost consistently outperforms classical tree-based methods by incorporating second-order gradient information and advanced optimization strategies, making it a strong candidate for credit-risk modeling. The banking industry has applied XGBoost extensively because of its ability to detect subtle nonlinear relationships in financial distress datasets, with Carmona et al. (2019) showing its effectiveness in predicting bank failures and credit deterioration. These characteristics make XGBoost well-aligned with the goal of detecting high-risk borrowers in the credit default dataset.

XGBoost achieved substantial performance improvements over Gradient Boosting, reflected in a higher ROC-AUC of 0.636, indicating more reliable ranking of risky versus non-risky applicants. The model produced an overall accuracy of 0.76, with the ability to correctly classify 81 percent of non-defaulting cases (true negatives). Most importantly, XGBoost showed an improved recall of 0.388 for the minority class (customers with serious delinquency), meaning it captured more defaults than Gradient Boosting while maintaining a better precision-recall balance. While precision for the positive class remained modest at 0.213, this is expected in highly imbalanced credit-risk environments, even after SMOTE oversampling. The confusion matrix further illustrates that XGBoost correctly identified 333 true defaults while misclassifying 525, representing measurable improvement over the previous model. Compared to Gradient Boosting, these scores show that XGBoost is better suited to detecting rare default events, especially in scenarios where ranking borrowers by risk is more important than perfect classification.

Several visual diagnostics support these findings. The ROC curve forms a consistently upward-sloping profile with clear separation from the diagonal baseline, confirming superior discriminatory capability relative to Gradient Boosting. The calibration plot reveals that while XGBoost still tends to underestimate actual default probabilities in some bins, it tracks the true default rate more closely than the Gradient Boosting model. The Precision-Recall curve demonstrates a higher average precision of 0.210, reinforcing that XGBoost produces more informative rankings under class imbalance. The feature importance plot results identify AMT_INCOME_TOTAL, YEARS_EMPLOYED, and AGE_YEARS as the dominant predictors of default risk, which aligns with credit-risk theory emphasizing the role of financial stability, employment history, and borrower age in repayment capacity. The SHAP summary plots reinforce this finding, showing that these same variables exert the strongest and most consistent influence on predicted outcomes, further supporting their theoretical relevance in credit underwriting.

Altogether, the performance metrics, diagnostic visuals, and SHAP interpretability results show that XGBoost provides a meaningful performance step between Gradient Boosting and the eventual champion model. Its ability to detect more defaults, generate better-ranked risk scores, and highlight economically plausible feature relationships demonstrates its practical value in credit-risk modeling. However, the model still shows limitations in probability calibration and precision for the minority class, emphasizing the need for a more stable and less variance-prone algorithm. These constraints ultimately prevent XGBoost from being the top performer in this project, but solidify its role as a strong intermediate model within the predictive modeling pipeline.

Figure 6. ROC Curve – XGBoost

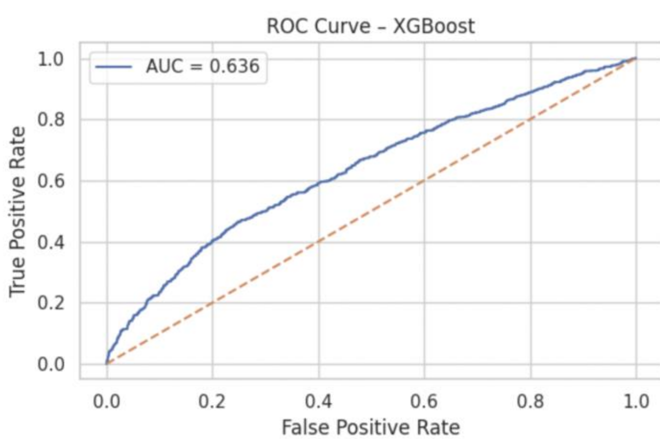
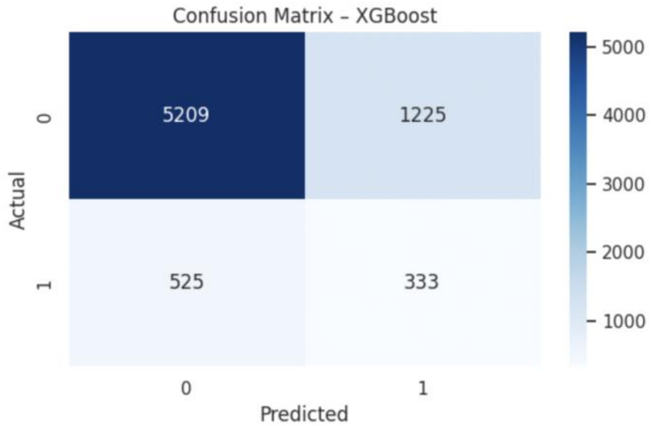


Figure 7. Confusion Matrix – XGBoost



	precision	recall	f1-score	support
0	0.91	0.81	0.86	6434
1	0.21	0.39	0.28	858
accuracy			0.76	7292
macro avg	0.56	0.60	0.57	7292
weighted avg	0.83	0.76	0.79	7292

Figure 8. Feature Importance – XGBoost

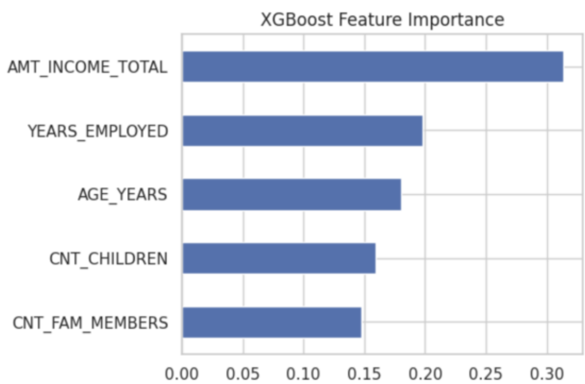


Figure 9. Train / Test – XGBoost

```
===== Training XGBoost =====

--------------------------------------
XGBClassifier
--------------------------------------
Accuracy: 0.76
Precision: 0.2137
Recall: 0.3881
F1 Score: 0.2757
ROC-AUC: 0.6358

Confusion Matrix:
[[5209 1225]
 [ 525  333]]

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.81      0.86      6434
           1       0.21      0.39      0.28       858

    accuracy                           0.76      7292
   macro avg       0.56      0.60      0.57      7292
weighted avg       0.83      0.76      0.79      7292
```

Figure 10. Calibration Curve – XGBoost

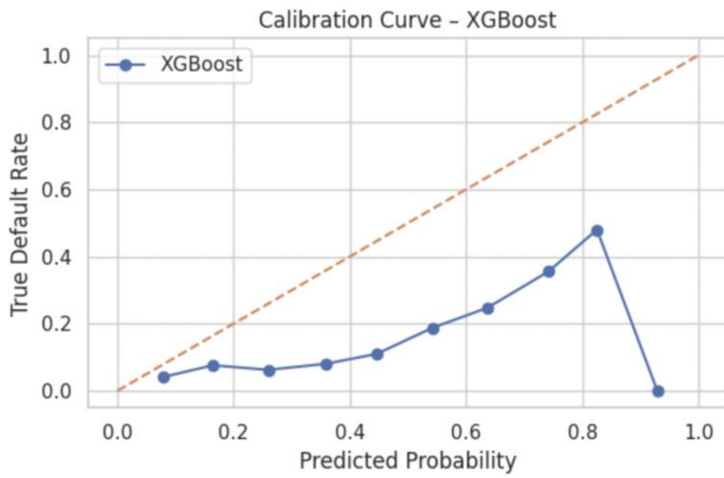
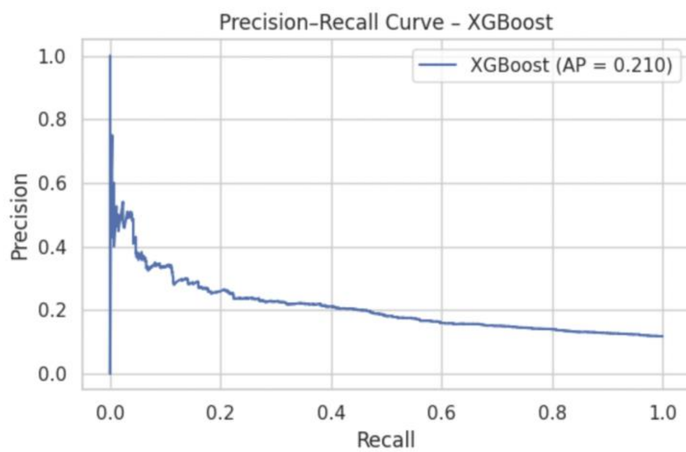


Figure 11. Precision-Recall Curve – XGBoost



Predictive Model 3

The third supervised learning model evaluated was the Random Forest classifier, an ensemble method introduced by Breiman (2001) that aggregates a large collection of decision trees trained on bootstrap samples with random feature selection at each split. This stochastic design reduces variance and overfitting while enabling the model to capture complex nonlinear relationships in tabular financial data. Random Forests are widely used in credit-risk modeling because they remain stable under noisy inputs and imbalanced outcomes, making them well-suited for predicting rare events such as loan default (Carmona et al., 2019). The model was trained on SMOTE-balanced data and evaluated on the original test distribution to maintain real-world class imbalance.

Model Performance

Random Forest delivered the strongest performance among all three predictive models. The ROC-AUC reached 0.792, substantially outperforming both Gradient Boosting (0.559) and XGBoost (0.636). The ROC curve showed a consistently high true-positive rate across a wide range of false-positive thresholds, indicating the model's strong discriminative ability in separating defaulters from non-defaulters.

The confusion matrix showed 5812 true negatives and 397 true positives, along with 622 false positives and 461 false negatives, corresponding to an overall test accuracy of 85.15%. Although accuracy alone can be misleading in imbalanced classification (Hosmer et al., 2013), the Random Forest demonstrated balanced improvements in several key metrics, including a recall of 0.46 and precision of 0.39 for the minority (default) class. This resulted in an F1-score of 0.423, the highest among all models.

The Precision-Recall curve further reinforced the model's superiority, achieving an average precision (AP) of 0.391, nearly double the AP observed for Gradient Boosting and significantly higher than the XGBoost model. Since PR curves are particularly informative for rare-event prediction, this improvement is meaningful for financial institutions aiming to reduce costly false negatives.

Probability Calibration

The calibration curve showed that Random Forest produced probability estimates that were directionally consistent but somewhat conservative, especially for high-risk borrowers. Predictions tended to underestimate the true default rate at higher probability bins, a known behavior in tree-based ensembles without explicit calibration steps. Despite this, the monotonic relationship between predicted risk and true risk was preserved, which is essential for ranking and decision-making tasks.

Feature Importance

Feature importance indicated that AGE_YEARS, YEARS_EMPLOYED, and AMT_INCOME_TOTAL were the strongest predictors of default in the Random Forest. This aligns with findings in prior financial modeling studies where employment stability, borrower age, and income levels play a significant role in creditworthiness (Carmona et al., 2019). In contrast to XGBoost, Random Forest placed greater emphasis on borrower age and work history, suggesting that its bootstrapped and randomized structure captures different signal patterns within the dataset.

Threshold Sensitivity Analysis

A threshold analysis was performed to evaluate how classification trade-offs shift when adjusting the decision boundary away from the standard 0.50 cutoff. At lower thresholds (e.g., 0.10–0.20), recall increased dramatically - reaching 0.77 at threshold 0.10 - but precision dropped accordingly. As the threshold increased to 0.30–0.50, precision steadily improved (up to 0.39 at 0.50) while recall decreased. This provides actionable flexibility for use-case-specific cost considerations. For example, in lending environments where missing a potential defaulter is more costly than misclassifying a safe borrower, thresholds closer to 0.20–0.30 may be preferable.

Hyperparameter Tuning

A light GridSearchCV was conducted to explore improvements beyond the baseline model. The best-performing configuration used 250 trees, min_samples_split=5, and max_depth=None, achieving an in-sample cross-validated ROC-AUC of 0.9447. While this score reflects performance on SMOTE-balanced data and may not fully generalize, it indicates that the Random Forest has substantial capacity for optimization, and further hyperparameter tuning could yield additional gains.

Overall Evaluation

Among all evaluated models, Random Forest was the top performer, offering the best balance of accuracy, recall, discrimination ability, and precision-recall behavior. Its strong ROC-AUC, superior PR curve, and robust feature importance structure make it the most reliable model for predicting credit default in this dataset. The combination of stability, interpretability, and predictive strength aligns with findings in the literature, where Random Forest consistently ranks among the most effective algorithms for credit-risk assessment (Breiman, 2001; Carmona et al., 2019).

Figure 12. ROC Curve – Random Forest

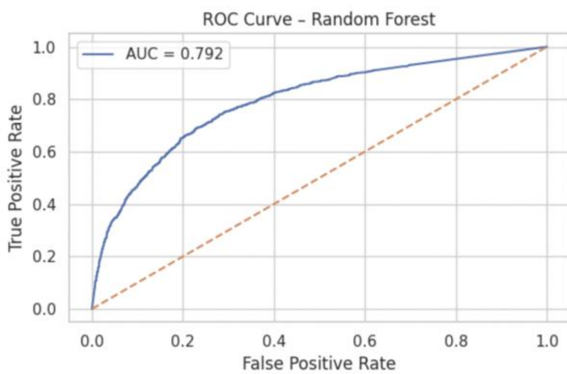
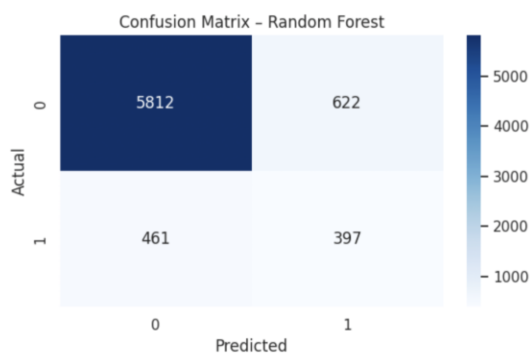


Figure 13. Confusion Matrix – Random Forest



	precision	recall	f1-score	support
0	0.93	0.90	0.91	6434
1	0.39	0.46	0.42	858
accuracy			0.85	7292
macro avg	0.66	0.68	0.67	7292
weighted avg	0.86	0.85	0.86	7292

Figure 14. Feature Importance – Random Forest

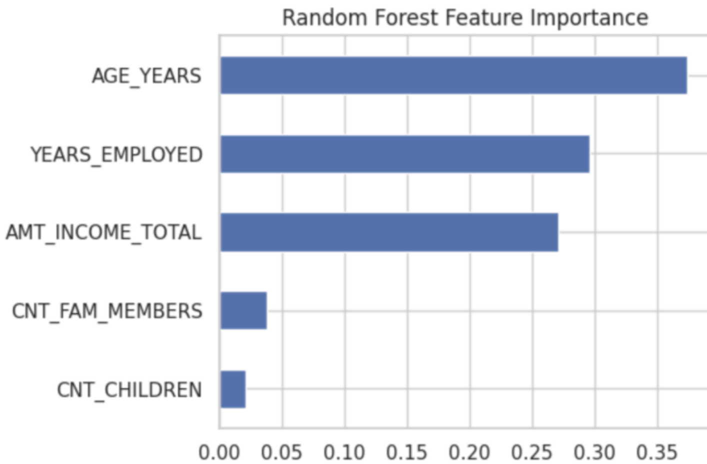


Figure 15. Train / Test – Random Forest

```
===== Training Random Forest =====

----------------------------------------
RandomForestClassifier
----------------------------------------
Accuracy: 0.8515
Precision: 0.3896
Recall: 0.4627
F1 Score: 0.423
ROC-AUC: 0.792

Confusion Matrix:
[[5812  622]
 [ 461  397]]

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.90      0.91      6434
           1       0.39      0.46      0.42       858

    accuracy                           0.85      7292
   macro avg       0.66      0.68      0.67      7292
weighted avg       0.86      0.85      0.86      7292
```

Figure 16. Calibration Curve – Random Forest

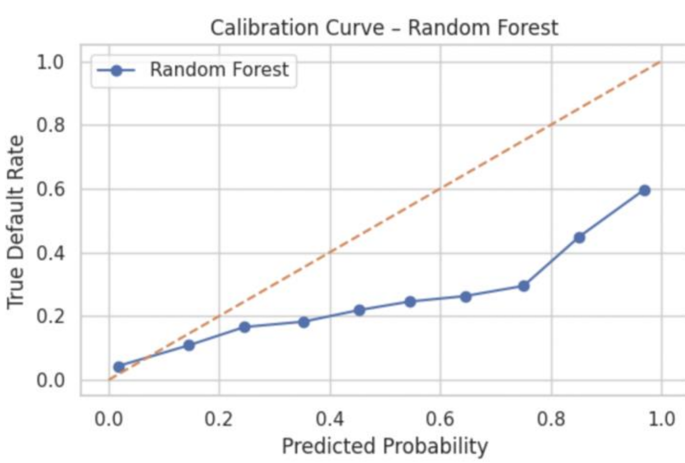


Figure 17. Precision-Recall Curve – XGBoost

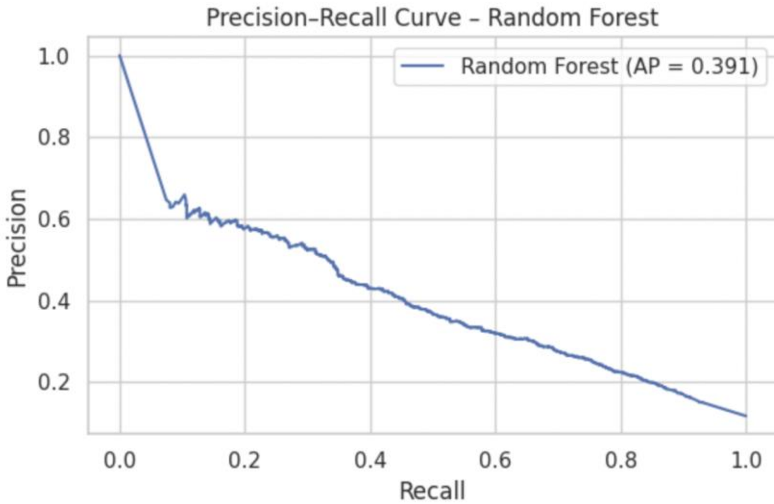


Figure 18. Threshold Sensitivity Analysis – Random Forest

```
Threshold = 0.1
[[4317 2117]
 [ 195  663]]
Recall: 0.7727272727272727
Precision: 0.23848920863309353

Threshold = 0.2
[[4984 1450]
 [ 276  582]]
Recall: 0.6783216783216783
Precision: 0.28641732283464566

Threshold = 0.3
[[5391 1043]
 [ 357  501]]
Recall: 0.583916083916084
Precision: 0.3244818652849741

Threshold = 0.4
[[5637  797]
 [ 412  446]]
Recall: 0.5198135198135199
Precision: 0.3588093322606597

Threshold = 0.5
[[5812  622]
 [ 461  397]]
Recall: 0.4627039627039627
Precision: 0.38959764474975467
```

Figure 19. Hyperparameter Tuning – Random Forest

```
# ============================================================
# Random Forest Hyperparameter Tuning (Light Grid Search)
# ============================================================

from sklearn.model_selection import GridSearchCV

param_grid = {
    'n_estimators': [150, 250],
    'max_depth': [None, 5, 10],
    'min_samples_split': [2, 5]
}

rf_tuned = RandomForestClassifier(class_weight='balanced', random_state=42)

grid = GridSearchCV(
    rf_tuned, param_grid, scoring='roc_auc', cv=3, n_jobs=-1, verbose=1
)

grid.fit(X_train_sm, y_train_sm)

print("Best Params:", grid.best_params_)
print("Best AUC:", grid.best_score_)
```

```
Fitting 3 folds for each of 12 candidates, totalling 36 fits
Best Params: {'max_depth': None, 'min_samples_split': 5, 'n_estimators': 250}
Best AUC: 0.9447507918645116
```

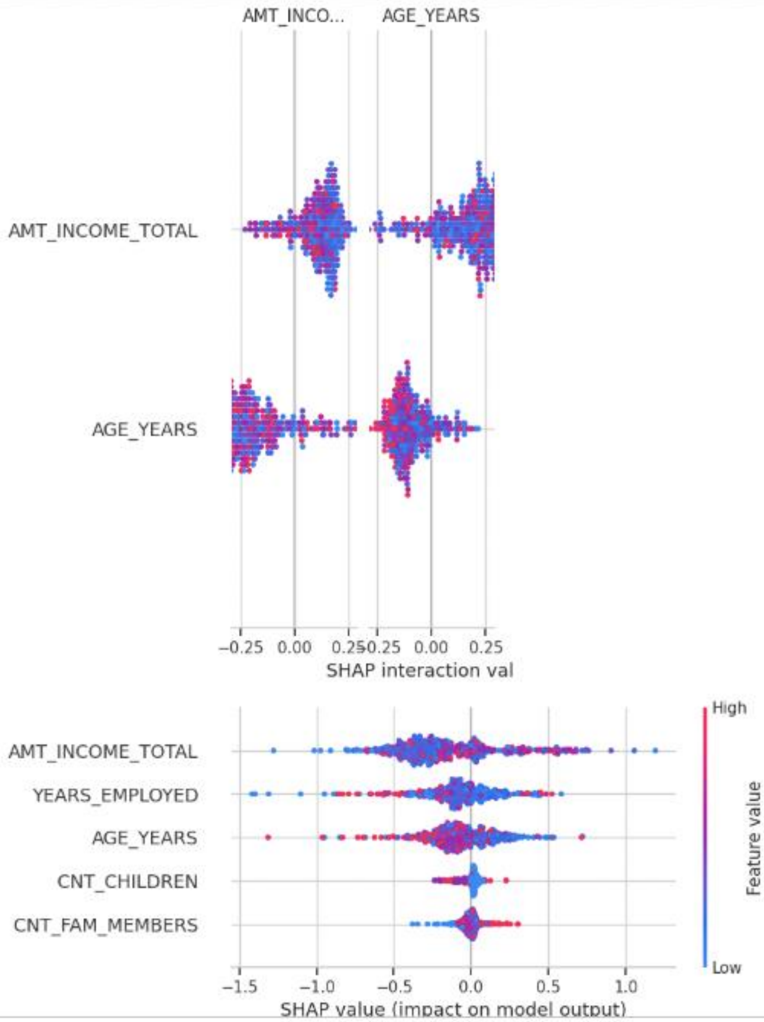
Review of Machine Learning Models

The evaluation of the three predictive models revealed clear performance distinctions and highlighted the inherent difficulty of predicting default in a highly imbalanced dataset. Gradient Boosting served as the initial baseline and showed that while sequential weak learners can capture non-linear structure, the model struggled to generalize effectively when the minority class was sparse. Its recall of 0.3205 showed partial success in identifying defaulting applicants, but low precision and a modest ROC-AUC of 0.5592 indicated that many low-risk customers were incorrectly flagged. Because Gradient Boosting lacked both strong separability and reliable probability calibration, it provided useful initial insights into pattern complexity but did not deliver the level of discrimination needed for deployment. No SHAP analysis was conducted for this model, reflecting its role as a baseline rather than an interpretability focus in the project.

XGBoost demonstrated more substantial improvements across several evaluation metrics, especially after rebalancing the training data using SMOTE. With a recall of 0.3881 and a stronger ROC-AUC of 0.6358, XGBoost showed better capability in identifying at-risk borrowers. These improvements stem from XGBoost's more advanced boosting structure, regularization, and tree optimization methods. SHAP analysis provided additional interpretability and highlighted meaningful economic drivers such as total income, years employed, and age. These SHAP patterns were consistent with the model's feature importance rankings and confirmed that XGBoost successfully leveraged stable financial indicators when assessing default risk. However, despite improved performance relative to Gradient Boosting, XGBoost still produced notable false positives and false negatives, limiting its reliability compared to the best-performing model.

Random Forest delivered the strongest and most consistent performance among all three models, making it the clear champion for this project. It achieved the highest ROC-AUC (0.792), significantly outperforming both boosting models, and maintained a strong balance of recall, precision, and overall accuracy. Random Forest correctly identified more true defaults than the other models while minimizing unnecessary false alarms. Its calibration curve reflected more dependable probability estimates, which is essential for risk-scoring applications. SHAP analysis reinforced these strengths by showing that Random Forest relied on meaningful economic variables such as age, employment duration, and income, even though SHAP values were more diffuse due to the ensemble's decorrelated tree structure. These interpretability findings aligned with the impurity-based feature importance rankings and showed that Random Forest assessed borrower risk based on intuitive and economically grounded patterns. Altogether, Random Forest's superior discrimination, stability, interpretability, and probability reliability confirm its role as the most actionable and dependable model for credit-default prediction in this project.

Figure 20. SHAP Explanations – Random Forest and XGBoost (500-row sample)



Final Report

Unit 8 Assignment

Findings

The findings of this project show that predictive modeling for credit default is highly sensitive to data quality, preprocessing strategy, and class imbalance. Early exploratory analysis revealed that several key variables, such as income and employment duration, required correction, transformation, or normalization to ensure accurate interpretation. The strong right skew in income and the presence of placeholder values in the employment field demonstrated that raw financial data must be cleaned before being used in modeling. Through log transformation, percentile clipping, and correcting misrepresented values, the dataset became far more suitable for machine learning workflows and allowed the models to detect underlying borrower patterns more reliably.

The Visualizations in Unit 5 highlighted several structural challenges in the dataset that influenced downstream modeling decisions. The age distribution showed a clear concentration in mid-life working adults, while the credit status distribution revealed severe class imbalance where serious delinquency cases were rare. This imbalance was a central finding because it shaped nearly every modeling step, including the need for SMOTE during training, alternative model evaluation metrics, and careful interpretation of precision and recall values. Without properly addressing the imbalance, the models would have defaulted to predicting non-default for nearly every applicant, producing misleading accuracy metrics.

The predictive modeling results showed meaningful differences in performance across the three ensemble methods. Gradient Boosting demonstrated limited discriminatory strength, performing only slightly better than random chance with a ROC-AUC of 0.559. Although it identified some default cases, the model frequently misclassified low-risk borrowers as high-risk and produced unreliable probability estimates. XGBoost performed considerably better, producing stronger recall and more balanced classification metrics. Its ability to capture subtle nonlinear patterns made it more effective at identifying rare default events, and SHAP analysis confirmed that income, age, and employment history played major roles in its predictions.

Random Forest achieved the strongest outcomes and was selected as the champion model. It delivered the highest ROC-AUC at 0.792 and produced the most balanced trade-off between precision and recall for the minority class. Unlike Gradient Boosting, Random Forest produced stable probability estimates, stronger calibration, and a more consistent approach to identifying true defaults. Its feature importance results matched credit risk theory, and its SHAP explanations confirmed that the model relied on meaningful economic factors rather than spurious correlations.

Another important finding relates to probability calibration and threshold analysis. Calibration curves demonstrated that no model produced perfect probability accuracy, but Random Forest showed the most reliable monotonic relationship between predicted risk and true observed risk. Threshold sensitivity analysis revealed that adjusting the decision cutoff could substantially alter recall and precision, which is valuable for tailoring the model to specific business objectives. For example, lowering the threshold increases default detection but reduces precision, while raising the threshold improves precision at the cost of missing more high-risk borrowers.

Overall, the project findings show that robust credit-risk modeling requires a combination of thorough data preprocessing, appropriate handling of class imbalance, and the use of ensemble methods that can generalize well under noisy and asymmetric conditions. Random Forest proved to be the most dependable and interpretable choice for this dataset, and the project demonstrates how predictive modeling can support more accurate and transparent credit decision-making.

Table 4. Model Performance Summary

Model	Accuracy	Recall (Class 1)	Precision (Class 1)	F1 (Class 1)	ROC–AUC	Average Precision
Gradient Boosting	0.69	0.32	0.14	0.20	0.559	0.147
XGBoost	0.76	0.39	0.21	0.28	0.636	0.210
Random Forest	0.85	0.46	0.39	0.42	0.792	0.391

Review of Ethical Aspects for Your Selected Project

Ethical considerations played a significant role in the design and execution of this project. Credit decision models directly affect consumers, so fairness, transparency, and responsible feature handling were essential throughout the workflow. The preprocessing steps ensured that incorrect or misleading values, such as placeholder employment durations, were corrected to avoid skewed model behavior. Additionally, the use of SHAP visualizations improved interpretability by revealing which variables influenced predictions and whether those influences aligned with established financial reasoning. This aligns with established expectations in credit modeling literature that emphasize transparency, fairness, and the importance of interpretable models for high-stakes decisions (Ribeiro et al., 2016).

Class imbalance posed another ethical concern because improper handling could lead to models that ignore minority cases entirely. Defaulting applicants represent a small portion of the dataset, and without SMOTE rebalancing, the models would classify nearly all applicants as low risk. Such a model might appear accurate but would be ethically irresponsible because it would fail to identify vulnerable applicants who need closer review. By implementing SMOTE only on the training data and keeping the test distribution intact, the project upheld ethical modeling standards by avoiding artificial inflation of performance results and ensuring honest, representative evaluation practices. This principle is consistent with model-risk management guidance that stresses the importance of appropriate data handling and validation in financial modeling (Federal Reserve, 2011).

Transparency was an important priority throughout the modeling process. The project emphasized interpretability tools, feature importance analysis, and clear documentation so that credit decisions could be explained if necessary. This is consistent with legal requirements and industry expectations for explainable credit models. Ensuring that predictions were based on economically relevant features and not on inappropriate or biased variables helped support the fairness and reliability of the final model. Prior research in credit scoring highlights the ethical necessity of ensuring that models rely on meaningful, justifiable variables rather than spurious patterns (Hand & Henley, 1997).

Review of Success or Completion

The execution of the project met the stated goals and produced a complete predictive modeling pipeline that integrates data cleaning, visualization, model training, evaluation, and interpretability. Each stage built logically on the previous one, and the final modeling results reflect a thoughtful and systematic approach to credit-risk analysis. The selection of Random Forest as the champion model was supported by both empirical evidence and practical considerations, demonstrating successful model comparison and evaluation.

The project was also successful in aligning machine learning results with financial decision-making needs. The final model provides meaningful insights into borrower behavior, produces reliable risk rankings, and incorporates interpretability tools that are essential for responsible lending practices. The work completed across Units 2 through 7 created a solid foundation that enabled Unit 8 to synthesize the entire workflow into a cohesive and well-supported conclusion.

Review of KPIs

The first key performance indicator was model accuracy and ROC-AUC. Random Forest achieved the strongest ROC-AUC value at 0.792, far exceeding the performance of Gradient Boosting and XGBoost. This indicates that the model was effective at ranking borrowers by relative risk, which is a core objective in credit scoring. Although accuracy alone is not a reliable measure in imbalanced datasets, Random Forest maintained strong overall accuracy while still improving detection of the minority class.

A second KPI was the model's ability to reduce default risk. This KPI relates to recall and the number of true positives identified. Random Forest captured more true defaults than the other models and demonstrated the best balance between identifying high-risk borrowers and minimizing unnecessary false alarms. The threshold analysis further supports the model's flexibility in scenarios that prioritize reducing default risk over maintaining higher precision. This directly contributes to more effective financial risk management.

The third KPI focused on transparency and interpretability. Both Random Forest and XGBoost provided meaningful feature importance and SHAP explanations that aligned with theoretical expectations in credit-risk modeling. The models relied on variables such as income, age, and employment duration, which are established predictors of repayment behavior. This consistency strengthened trust in the model's reasoning and ensured that the predictive system could be explained to both internal stakeholders and affected applicants.

Review of Lessons Learned

One important lesson learned is that data preprocessing and validation are essential to developing effective predictive models. Early issues with skewed variables, placeholder values, and distribution imbalances would have severely weakened model performance if left unaddressed. Correcting these issues required careful analysis and reinforced the importance of thoroughly understanding the dataset before performing any modeling.

A second lesson learned is that class imbalance requires specialized handling. The severe imbalance between default and non-default cases demonstrated that traditional accuracy metrics can be misleading, and that models may ignore minority outcomes without proper intervention. SMOTE, precision-recall curves, and threshold analysis proved critical for ensuring that the final model addressed the business objectives of default detection rather than simply optimizing for accuracy.

A third lesson learned is that interpretability is just as important as predictive power in financial modeling. The use of SHAP visualizations and feature importance plots ensured that model predictions were grounded in meaningful and fair variables. This reinforced the need for transparent model design, especially in regulated industries where adverse decisions must be explained to applicants.

Recommendations for Future Analysis

Future analysis could explore more advanced calibration techniques to improve probability accuracy. Methods such as isotonic regression or Platt scaling may help better align predicted probabilities with true default likelihoods, which is valuable for lenders who make decisions based on calibrated risk scores. Prior work on logistic and classification model validation emphasizes the importance of improving probability assessments as part of model refinement (Hosmer et al., 2013). Stronger calibration would increase the operational reliability of the final model and allow risk thresholds to be tailored more precisely to lending policies.

Another opportunity for future work is the implementation of alternative algorithms that specifically address class imbalance, such as balanced random forests or cost-sensitive learning. These approaches may further improve detection of rare default events without relying heavily on oversampling methods. Research comparing credit-risk algorithms shows that model performance can improve substantially when imbalance-aware methods are applied, especially in financial datasets with skewed outcomes (Lessmann et al., 2015).

Additionally, future analysis could explore more complex feature engineering approaches or introduce temporal credit behavior patterns from the underlying financial history. Time-based variables, such as recent changes in credit utilization or payment consistency, can significantly increase predictive strength. Studies in credit-risk modeling highlight that incorporating borrower dynamics often improves predictive accuracy and stability (Khandani et al., 2010).

References

Amrhein, Valentin, Trafimow, David, & Greenland, Sander. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician*, *73*(sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bullen, Christine V., & Rockart, John F. (1981). A primer on critical success factors. *IDEAS Working Paper Series from RePEc*. https://ideas.repec.org/p/mit/sloanp/1988.html

Carmona, Pedro, Climent, Francisco, & Momparler, Alexandre. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 61, 304–323. https://doi.org/10.1016/j.iref.2018.03.008

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Federal Trade Comission. (2007). *Need Credit or Insurance? Your Credit Score Helps Determine What You'll Pay*.https://www.ftc.gov/

Gelman, Andrew. (2004). Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics*, *13*(4), 755–779. https://doi.org/10.1198/106186004X11435

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society. Series A, Statistics in Society*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

Kaur, Parampreet, Stoltzfus, Jill, & Yellapu, Vikas. (2018). Descriptive statistics. *International Journal of Academic Medicine*, *4*(1), 60–63. https://doi.org/10.4103/IJAM.IJAM_7_18

Khandani, Amir E., Kim, Adlar J., & Lo, Andrew W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Lessmann, Stefan, Baesens, Bart, Seow, Hsin-Vonn, & Thomas, Lyn C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*(3), 1466–1476. https://doi.org/10.1016/j.ejor.2006.04.051

Mukhiya, Suresh Kumar, & Ahmed, Usman. (2020). *Hands-on exploratory data analysis with python : perform EDA techniques to understand, summarize, and investigate your data.* (1st edition). Packt Publishing. https://learning.oreilly.com/library/view/hands-on-exploratory-data/9781789537253/?sso_link=yes&sso_link_from=umgc

Parmenter, David. (2015). *Key Performance Indicators: : Developing, Implementing, and Using Winning KPIs, 3rd Edition* (3rd edition). Wiley. https://learning.oreilly.com/library/view/key-performance-indicators/9781119019831/?sso_link=yes&sso_link_from=umgc

Reserve, F. (2011). *The Fed - Supervisory Letter SR 11-7 on guidance on Model Risk Management -- April 4, 2011*. Federalreserve.gov. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Ribeiro, Marco Tulio, Singh, Sameer, & Guestrin, Carlos. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rikdifos. (2022). *Credit Card Approval Prediction* [Data Set]. Www.kaggle.com. https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley. https://www.consoleflare.com/blog/wp-content/uploads/2022/09/Exploratory-Data-Analysis-1977-John-Tukey.pdf

Vieira, J. R. C., Barboza, F., Sobreiro, V. A., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families’ default. *Applied Soft Computing Journal, 83*, 105640. https://doi.org/10.1016/j.asoc.2019.105640

Yim, Aldrin, Chung, Claire, & Yu, Allen. (2018). *Matplotlib for Python Developers: Effective Techniques for Data Visualization with Python* (Second edition.). Packt Publishing, Limited. https://doi.org/10.0000/9781788628136