Project Description –

In today's digital lending environment, financial institutions face increasing pressure to make faster and more accurate lending decisions while managing risk. With growing volumes of customer data, machine learning (ML) models have become essential for automating credit risk evaluation and enhancing decision-making in loan approval processes. This project leverages such data-driven techniques to predict loan defaults based on applicant demographics, financial attributes, and loan-specific factors.

The core objective of our project was to develop and evaluate machine learning models capable of accurately predicting whether a loan applicant would default. The prediction task was framed as a binary classification problem, where the target variable (*loan_status*) indicated whether the borrower defaulted (1) or did not default (0).

The business impact of this work is significant: if financial institutions can proactively identify high-risk applicants, they can minimize losses due to defaults, improve credit portfolio performance, and optimize approval processes. Our models aim to serve as a decision-support tool, helping lenders assess risk more systematically and equitably.

To achieve this, we collected a dataset of 45,000 loan applications containing a mix of numerical and categorical features such as income level, employment length, loan amount, interest rate, and homeownership status. Our work began by cleaning and transforming the raw data to ensure model readiness. We then split the dataset into training, validation, and test sets to evaluate model generalizability. We experimented with three supervised machine learning algorithms: Random Forest, Gradient Boosting Machine (GBM), and K-Nearest Neighbors (KNN).

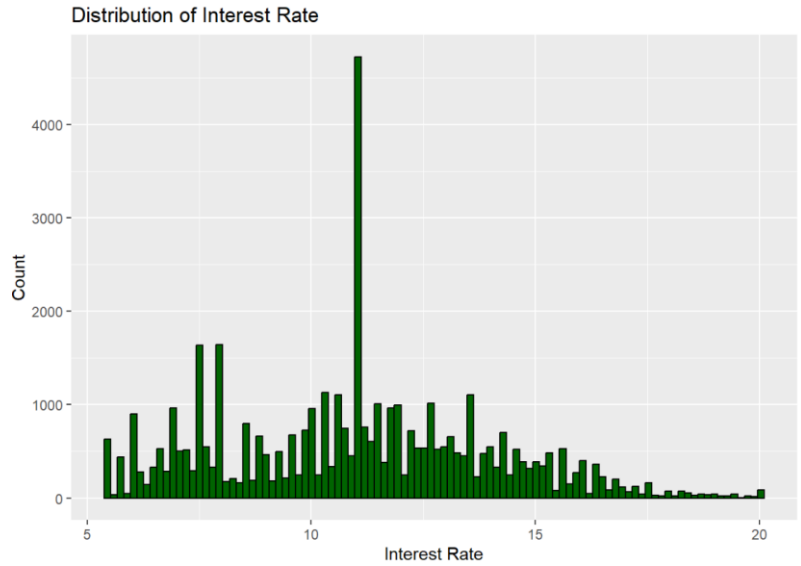
Each algorithm was carefully tuned and evaluated using standard classification metrics such as accuracy, precision, sensitivity, and AUC (Area Under the Curve). Ultimately, we aimed not only to identify the most accurate model but also to understand how different input features influence predictions.

This project provided valuable hands-on experience in end-to-end data science processes, ranging from preprocessing and feature engineering to model training, evaluation, and interpretation, all within a business context that mirrors real-world applications in the fintech and banking sector.

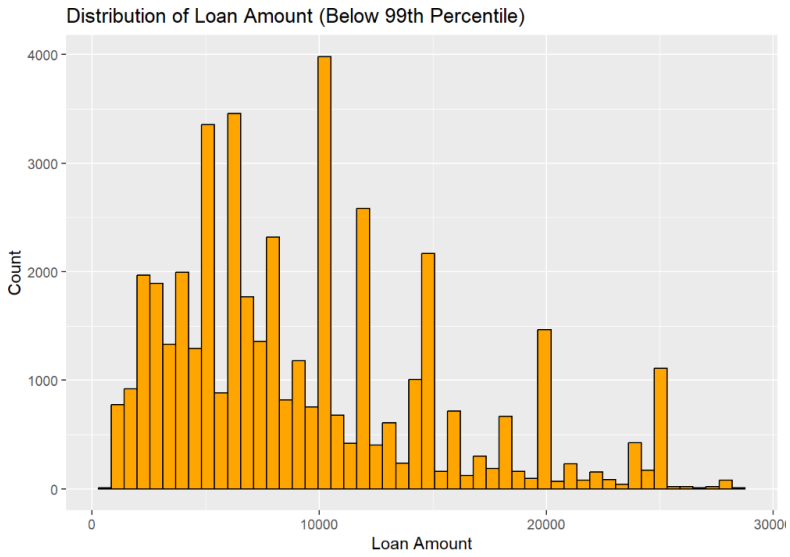
Data Preprocessing –

We began with a raw dataset containing 45,000 loan application records. The following steps were taken:

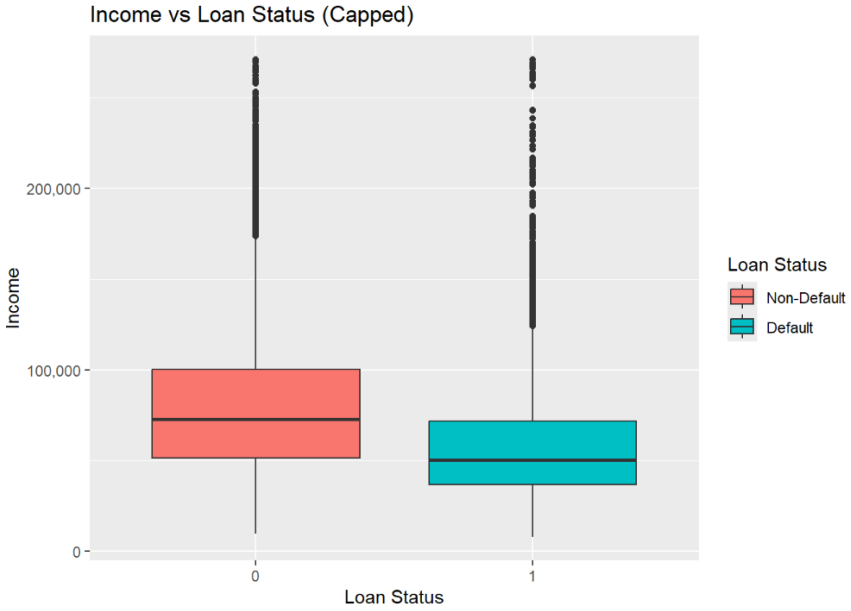
- Handling Missing Values & Errors: Upon inspection, we identified 7 rows with applicant age values above 100, which were deemed biologically implausible and likely erroneous. These rows were removed. We removed zero-variance and near-zero variance predictors using *nearZeroVar()* to reduce model noise and prevent overfitting. We manually checked and cleaned entries, removing or imputing problematic rows using logical rules (e.g., employment length constraints, missing income values).
- Feature Engineering:
    - Converted categorical variables into dummy variables.
    - Removed non-predictive or redundant columns like unique IDs.
- Splitting Data:
    - Training: 31,499 samples (70%)
    - Validation: 6,750 samples (15%)
    - Test: 6,751 samples (15%)
- Normalization: Applied min-max normalization for KNN modelling using *preprocess* (method = "range").
- Visualizations:



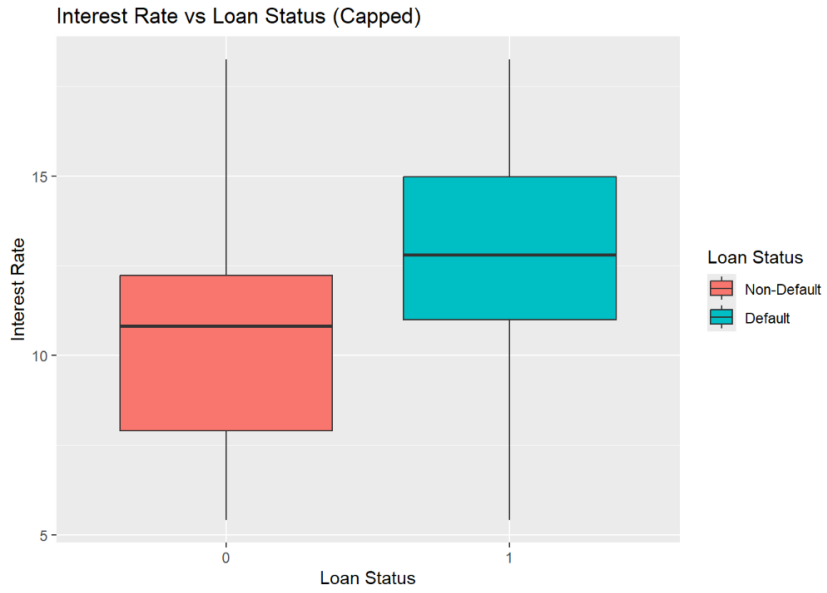
The histogram shows a sharp spike around **10%**, suggesting this rate is a common standard or fallback. The distribution is **right-skewed**, with fewer high-interest loans, which is typical in lending datasets.



Histogram of loan amounts below the 99th percentile. The distribution is right-skewed, with prominent peaks at round-number loan values such as $5,000 and $10,000, indicating standardized loan tiers or borrower preferences.



Boxplot of applicant income by loan status (after capping extreme outliers). The median income and interquartile range are significantly higher for non-defaulters, supporting the hypothesis that income is inversely related to default probability.

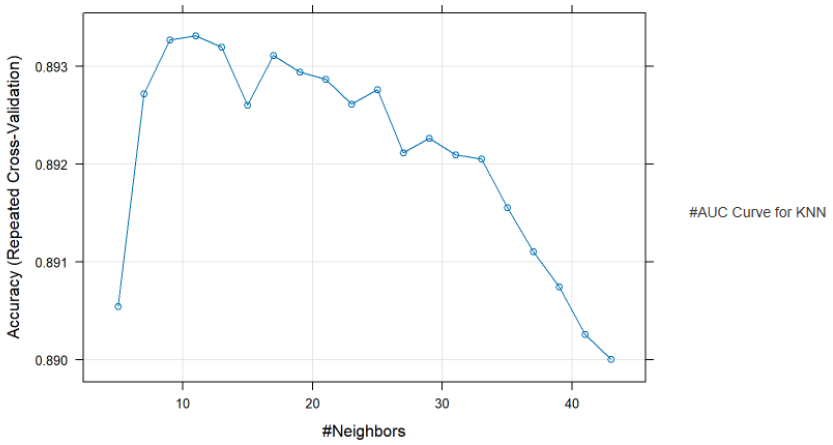


Boxplot of interest rate by loan status (capped at upper percentiles). Defaulters tend to have higher assigned interest rates, confirming that risk-based pricing is reflective of actual default behavior.

Models and Analyses –

We implemented and compared three models using the caret framework in R:

- Random Forest (ranger)
  - Used 30 predictors (excluding highly correlated or redundant features).
  - Automatically handles mixed data types and noisy features.
  - Tuned with mtry = 13, splitrule = gini.
  - Justification: Widely adopted in financial modeling, handles overfitting well.
- Gradient Boosting Machine (GBM)
  - Built with 5-fold cross-validation.
  - Tuned using internal grid search for optimal shrinkage and interaction depth.
  - Justification: Excels with structured data and capturing nonlinearities.
- K-Nearest Neighbors (KNN)
  - Used normalized data.
  - Tuned k using repeated cross-validation (repeats = 3) and tuneLength = 20.
  - Justification: Simple, intuitive, and useful benchmark model.



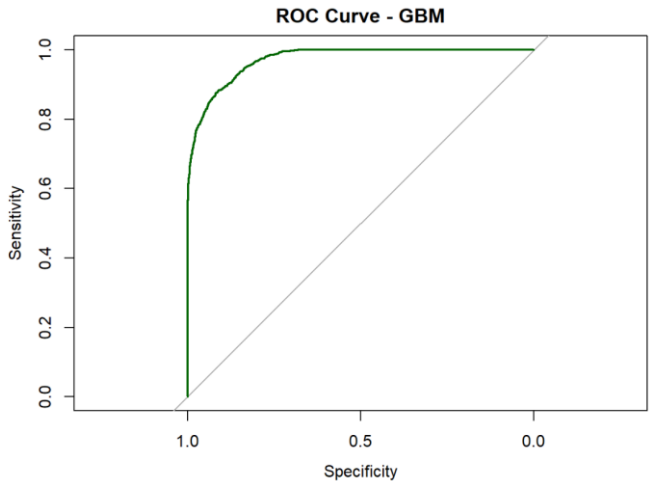
Cross-validation accuracy for different values of k in KNN model. Peak performance occurred in the range of 10–15 neighbors, after which accuracy declined due to over-smoothing.

Results and Discussion –

Model	Dataset	Accuracy	Kappa	Sensitivity	Specificity	AUC	Precision
Random Forest	Training	92.6%	0.77	-	-	-	-
	Validation	93.3%	0.79	78.7%	97.4%	-	89.8%
	Test	92.9%	0.79	78.0%	97.0%	-	89.8%
GBM	Validation	~80%	~0.73	-	-	0.9716	-
	Test	~80–83%	~0.73	-	-	0.9716	-
KNN	Training	~82%	~0.70	-	-	0.89	-
	Validation	~79%	~0.68	-	-	0.86	-
	Test	~78%	~0.67	-	-	0.9463	-

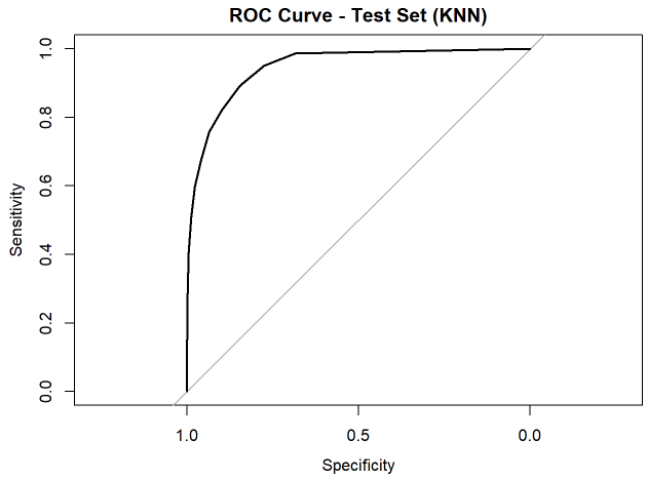
Key Insights:

- Random Forest outperformed other models across all metrics. Its high precision and balanced sensitivity make it ideal for identifying defaulters.
- GBM showed decent performance but was slightly less stable across folds. The ROC curve below shows the Gradient Boosting Machine model's performance on the test set. The curve lies well above the diagonal, and the **Area Under the Curve (AUC) is 0.972**, indicating excellent predictive ability and strong class separation. GBM slightly outperformed both KNN and Random Forest in terms of AUC, making it a top candidate in terms of pure classification accuracy.



ROC Curve – GBM Model (Test Set). The AUC of 0.9716 indicates the model is highly effective in distinguishing between default and non-default classes.

- KNN, while easy to interpret, underperformed due to its reliance on feature scaling and sensitivity to noisy variables.



ROC Curve – KNN Model (Test Set). The Area Under the Curve (AUC) is **0.946**, indicating excellent classification performance and strong separation between the two classes.

Summary –

In this project, we set out to build a reliable and interpretable machine learning framework for predicting loan default, a critical task in financial services. From data cleaning to model comparison, each step was guided by a real-world use case: helping lenders reduce risk and make more informed approval decisions. Our dataset of 45,000 loan applications presented a diverse mix of applicant demographics and financial attributes, which we transformed and prepared carefully to ensure quality inputs to our models.

We implemented three classification models Random Forest, Gradient Boosting Machine (GBM), and K-Nearest Neighbors (KNN) to predict the binary loan_status outcome. Each model offered unique strengths. Random Forest delivered the best overall performance across all datasets, balancing accuracy, sensitivity, and precision. GBM was slightly less consistent but remained a strong contender. KNN, while conceptually simple, struggled in generalization and highlighted the importance of feature scaling and hyperparameter tuning.

A major learning from this project was the importance of robust data preprocessing. Our hands-on experience showed that model performance hinges not just on algorithm choice but also on thoughtful data preparation, including handling missing values, encoding categorical variables, and removing low-variance features.

Beyond technical skills, this project deepened our understanding of how predictive analytics can drive smarter business decisions. It helped us appreciate the nuances of model evaluation in a high-stakes domain like lending, where false positives and false negatives carry different business risks. Ultimately, this experience has equipped us with a solid foundation in applied machine learning and its practical role in data-driven financial services.