K6 – MACHINE LEARNING

Proactively Managing Loan Risk

Leveraging Machine Learning to Reduce Late Payment

Presenter: Can Thi Hong Giang

06/2025

AGENDA

Predict Late Payment

01 PROJECT OBJECTIVE

02 CURRENT DATA STATUS

03 METHODOLOGY

04 MODEL IMPLEMENTATION & RESULTS

05 CONCLUSION & IMPROVEMENTS

The Challenge of Late Payments: Protecting Our Financial Health

Situation:

- ✓ **Brief overview:** Our institution processes a significant volume of loans (over 81,000 loans within the past 4 years).
- ✓ **Current state:** While most loans are paid on time, a subset of customers face challenges leading to late payments, which can escalate to defaults. Late or defaulted payments significantly affect cash flow and credit risk management.

Complication:

- ✓ **Urgency:** Late payments are early indicators of potential default, leading to increased collection efforts, strained customer relationships, and eventual financial losses if not managed.
- ✓ **Impact:** Accumulating late payments impact cash flow, increase operational costs for collections, and serve as a precursor to more significant portfolio-wide risk.
- ✓ **Why a solution is needed?**
    - Identifying *which specific customers* are likely to pay late allows for targeted, early intervention..
    - A reactive approach to late payments is less effective and more resource-intensive than proactive measures.
    - Key Question: How can we accurately classify which customers are at risk of late payment (Yes/No) to enable timely and effective mitigation strategies?

Resolution:

- ✓ Develop a Machine Learning model to proactively predict loans at risk of late payment, enabling better risk mitigation and resource allocation.

Understanding Our Data Landscape

Data Sources:

- ✓ Dataset from Kaggle

Data Volume & Scope:

- ✓ Total loans analyzed: 81,000+ of loan records over 4 years.
- ✓ Target variable: Loan_status
- ✓ Key variables considered: 22 features including applicant financial health, demographic details and loan characteristics.

Data Quality & Preprocessing:

- ✓ Key issues identified:
  - 5% missing values in ‘emp_length’.
  - 'emp_title' contains many unique and unstandardized values.
  - The income data is highly skewed and contain many outliers.
  - The datetime data is in the wrong format, and some columns need to be cleaned.
- ✓ Actions taken:
  - Imputation of missing values using median.
  - Process, embed, cluster data using KMeans.
  - Handle outliers and normalize the data using StandardScaler.
  - Process data and remove redundant columns.

Data Dictionary

Variable Name	Role (Feature/Target)	Type	Description
emp_length	Feature	Object	Length of employment (e.g., "10+ years").
home_ownership	Feature	Object	Type of home ownership (e.g., "MORTGAGE", "RENT", "OWN", "OTHER").
annual_inc	Feature	Float64	Annual income of the customer.
annual_inc_joint	Feature	Float64	Joint annual income (if applicable).
verification_status	Feature	Object	Status of income verification (e.g., "Verified", "Source Verified", "Not Verified").
avg_cur_bal	Feature	Float64	The average amount a customer owes across all their active credit accounts.
Tot_cur_bal	Feature	Float64	The total sum of all outstanding debts a borrower has right now.
loan_status	Target	Object	Loan status indicating whether the loan is late payment or not (class 1: yes, class 0: no).
loan_amount	Feature	Float64	Amount of the loan requested.
term	Feature	Object	Term of the loan (e.g., "36 months", "60 months").
int_rate	Feature	Float64	Interest rate of the loan.
installment	Feature	Float64	Monthly installment amount.
grade	Feature	Object	Credit grade assigned to the loan (e.g., "A", "B", "C", etc.).
issue_month	Feature	Int32	Month when the loan was issued.
issue_quarter	Feature	Int32	Quarter when the loan was issued.
issue_year_num	Feature	Int32	Year when the loan was issued.
pymnt_plan	Feature	Boolean	Indicates if there is a payment plan for the loan.
type	Feature	Object	Type of loan (e.g., "Individual", "direct pay", etc.).
purpose	Feature	Object	Purpose of the loan (e.g., "debt_consolidation", "credit_card", etc.).
subregion	Feature	Object	Subregion where the customer resides.
job_level	Feature	Object	Job level or position of the customer.
profession	Feature	Object	Profession of the customer.

How We Defined “Late Payment” and “On Time Payment” Loan

Define label classes based on risk of financial loss:

✓ **Late Payment (Target =1):**

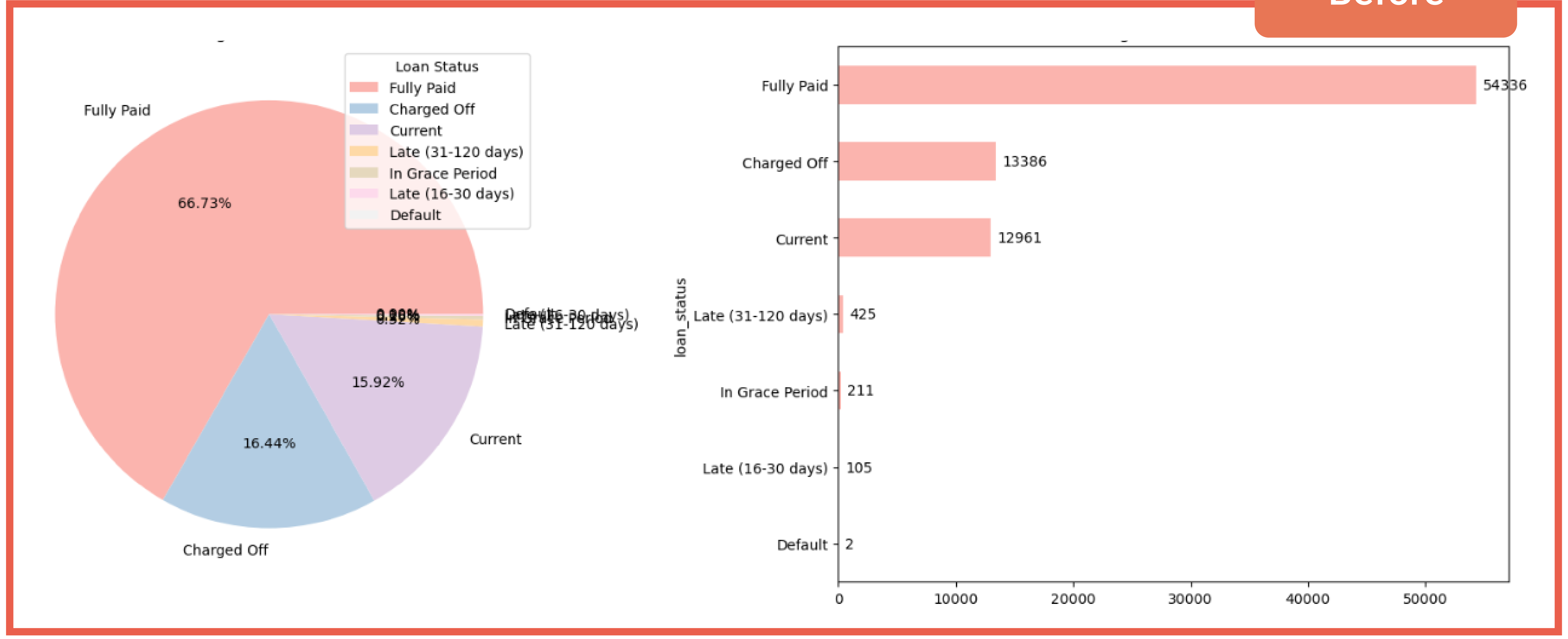
- Charged Off’
- 'Default’
- 'Late (16-30 days)’
- 'Late (31-120 days)’

✓ **On Time Payment (Target =0):**

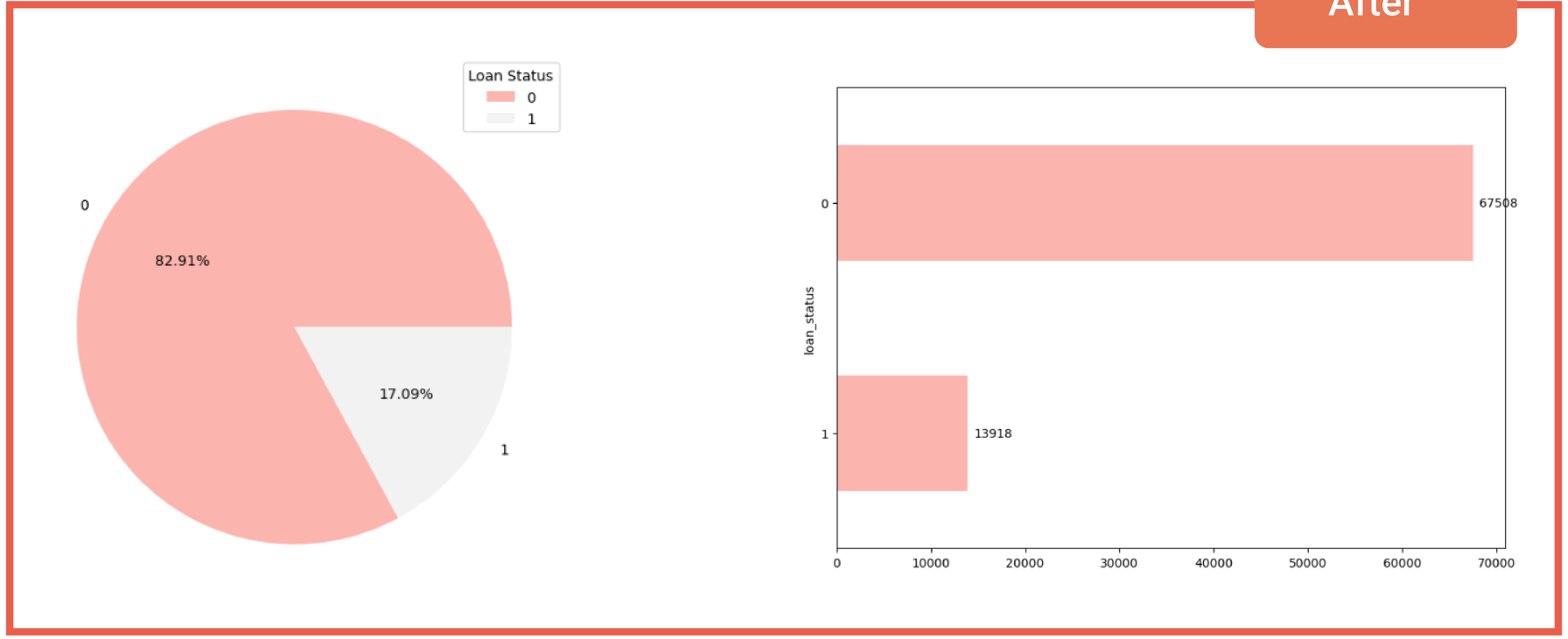
- 'Fully Paid’
- 'Current’
- 'In Grace Period’

➢ The dataset is imbalanced, with **17%** in class 1 and **83%** in class 0.

Before



After



Snapshot of Our Loan Portfolio

Categorical variables

	Cat-col	dtype	num_unique_value	values	num_null
0	home_ownership	object	6	[RENT, MORTGAGE, OWN, ANY, OTHER, NONE]	0
1	verification_status	object	3	[Verified, Not Verified, Source Verified]	0
2	term	object	2	[36 months, 60 months]	0
3	grade	object	7	[C, A, B, D, E, F, G]	0
4	pymnt_plan	bool	2	[False, True]	0
5	type	object	3	[INDIVIDUAL, JOINT, DIRECT_PAY]	0
6	purpose	object	13	[credit_card, home_improvement, debt_consolidation, other, vacation, major_purchase, medical, house, car, moving, small_business, renewable_energy, wedding]	0
7	subregion	object	9	[New England, Mountain, Middle Atlantic, West North Central, South Atlantic, East North Central, Pacific, East South Central, West South Central]	0
8	job_level	object	7	[No Information, Entry Level, Management, C-level, Individual Contributor, Mid-level, Director-level]	0
9	profession	object	26	[No Information, Healthcare/Medical, Manufacture/Distributor, Management/Specialist/Supervisor, Others, Sales/Marketing, Worker, Logistics/Delivery/Driver, Civil Servant, Mechanic/Maintenance, Security, Educator/Teaching, IT/Technician/Engineer, Admin/Assistant/Support/Services, Agent/Legal/Insuarance, Counselor/Therapist, Representative/Relations, Consultant, Convenience Services, Financial/Accounting/Analyst, Food and Beverage, Operations, Production/Assembler, Clerk, Coordinator, Human Resources]	0
10	have_inc_joint	object	2	[no, yes]	0

Numerical variables

	emp_length	avg_cur_bal	Tot_cur_bal	loan_amount	int_rate	installment	issue_month	issue_quarter	issue_year_num	loan_age_days	total_inc
count	77226.000000	81426.000000	8.142600e+04	81426.000000	81426.000000	81426.000000	81426.000000	81426.000000	81426.000000	81426.000000	8.142600e+04
mean	6.845486	13271.463169	1.390987e+05	15014.095621	0.133168	443.736108	7.099379	2.730565	2014.285793	439.965723	7.566194e+04
std	4.394070	15957.313943	1.535651e+05	8439.872249	0.044229	244.327605	3.394966	1.109315	0.851983	316.649433	6.737378e+04
min	0.500000	0.000000	0.000000e+00	1000.000000	0.053200	29.520000	1.000000	1.000000	2012.000000	30.000000	4.000000e+03
25%	3.000000	3138.000000	2.985850e+04	8425.000000	0.099900	267.209990	4.000000	2.000000	2014.000000	183.000000	4.600000e+04
50%	7.000000	7352.000000	8.059050e+04	13575.000000	0.129900	388.109990	7.000000	3.000000	2015.000000	364.000000	6.500000e+04
75%	12.000000	18431.750000	2.078412e+05	20000.000000	0.162900	580.349980	10.000000	4.000000	2015.000000	670.000000	9.000000e+04
max	12.000000	555925.000000	4.447397e+06	35000.000000	0.289900	1424.569900	12.000000	4.000000	2015.000000	1247.000000	8.900060e+06

Predicting Loan Risk: A Data-Driven Approach

Situation:

✓ Due to the increasing complexity of customer financial profiles and market dynamics, early risk detection has become essential.

Complication:

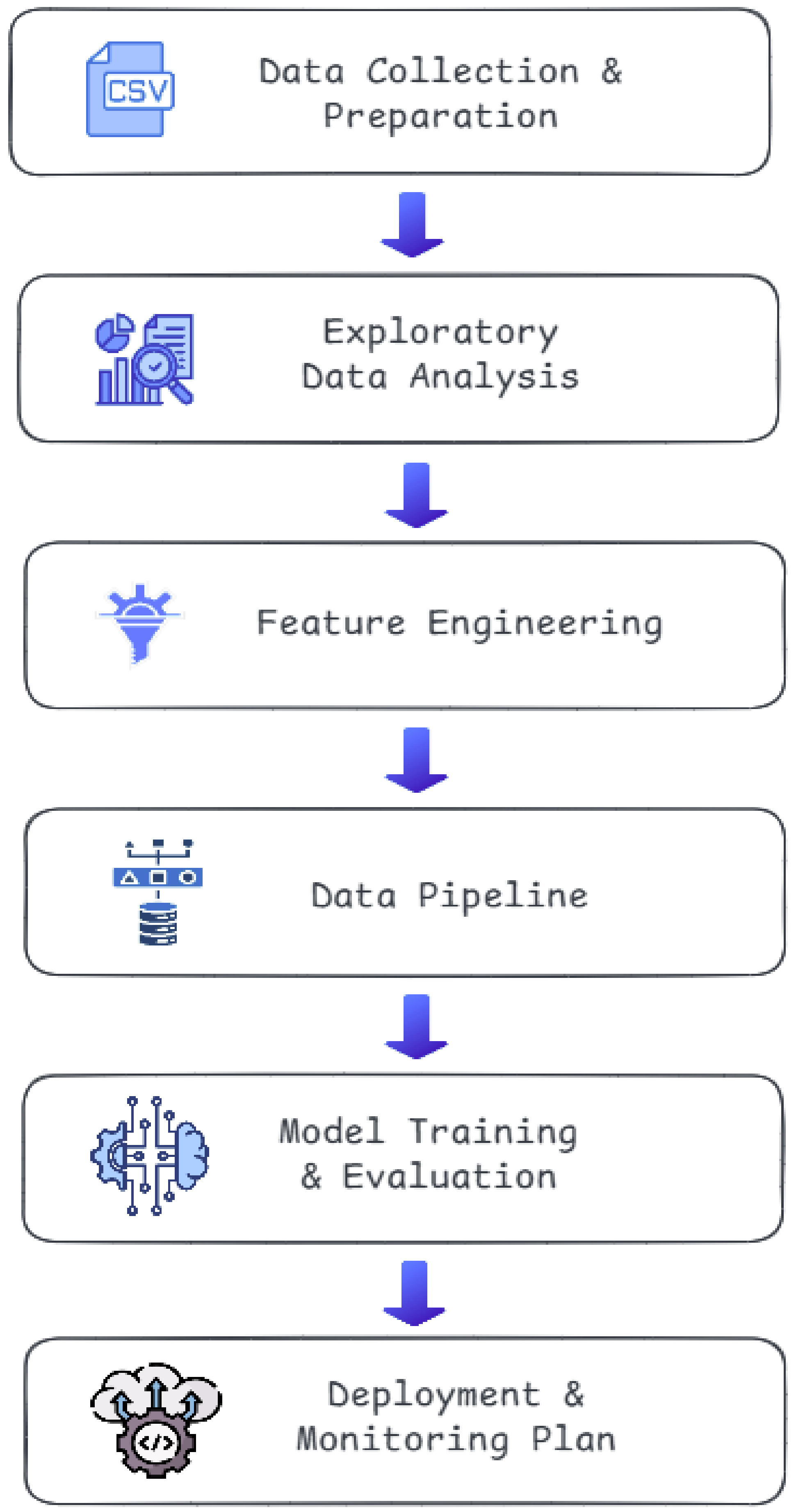
✓ Manually identifying loans at risk of late payments is inefficient and often reactive, leading to missed opportunities for intervention and increased financial losses. We need a systematic and accurate method to flag these loans early.

Resolution:

✓ Develop a robust model that accurately classifies customers into two groups: those likely to make a late payment ('Yes - At Risk') and those not likely ('No - Not At Risk').

Estimated Result:

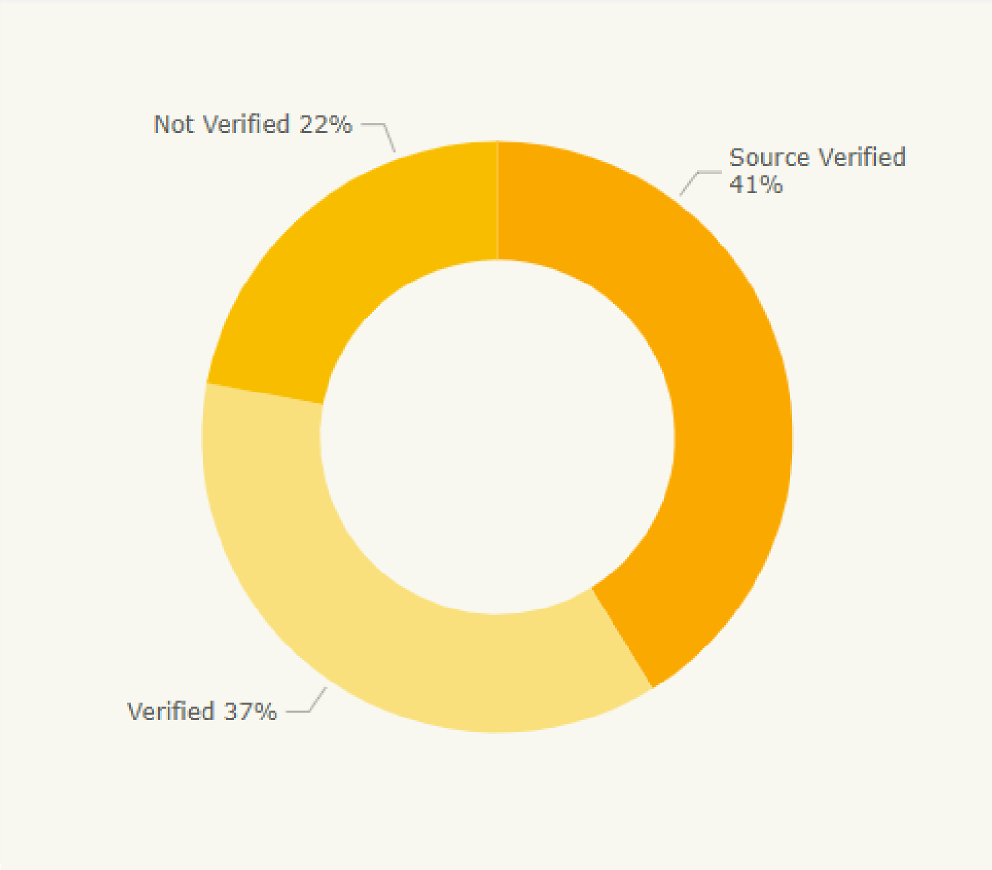
✓ The model is expected to achieve a recall of 0.7–0.8, significantly improving the ability to identify risky loans and enabling proactive intervention.

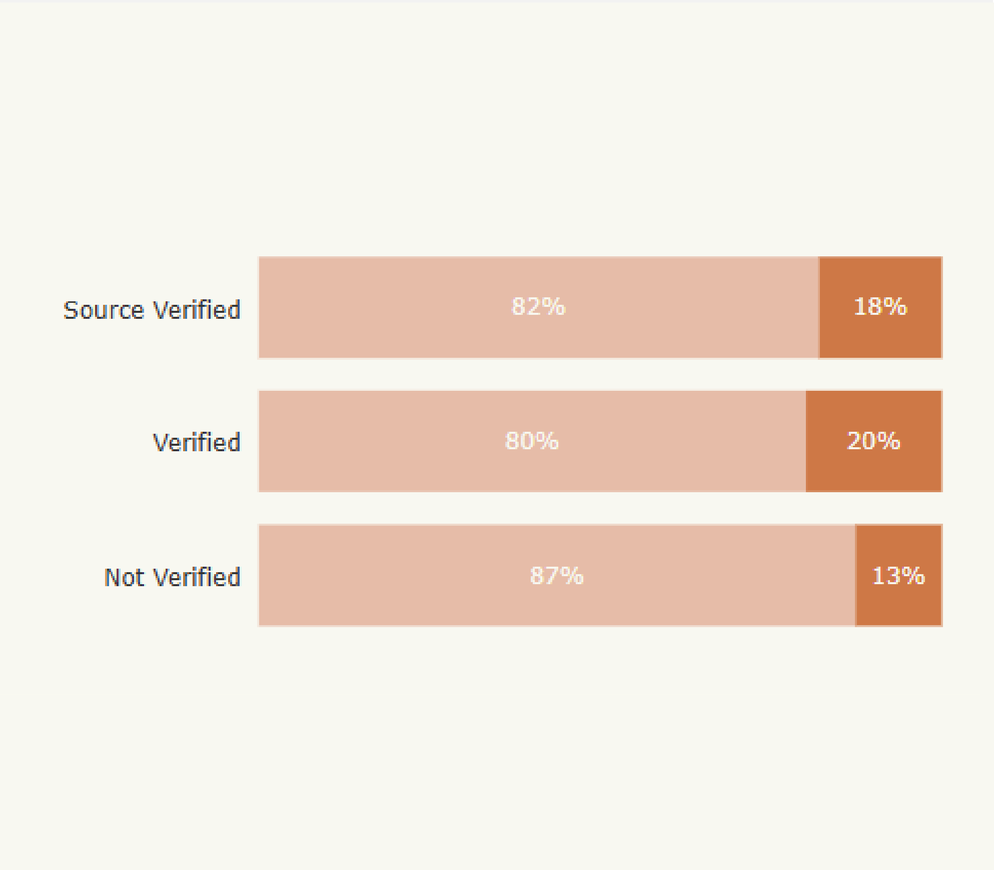


Distribution of Key Categorical Features

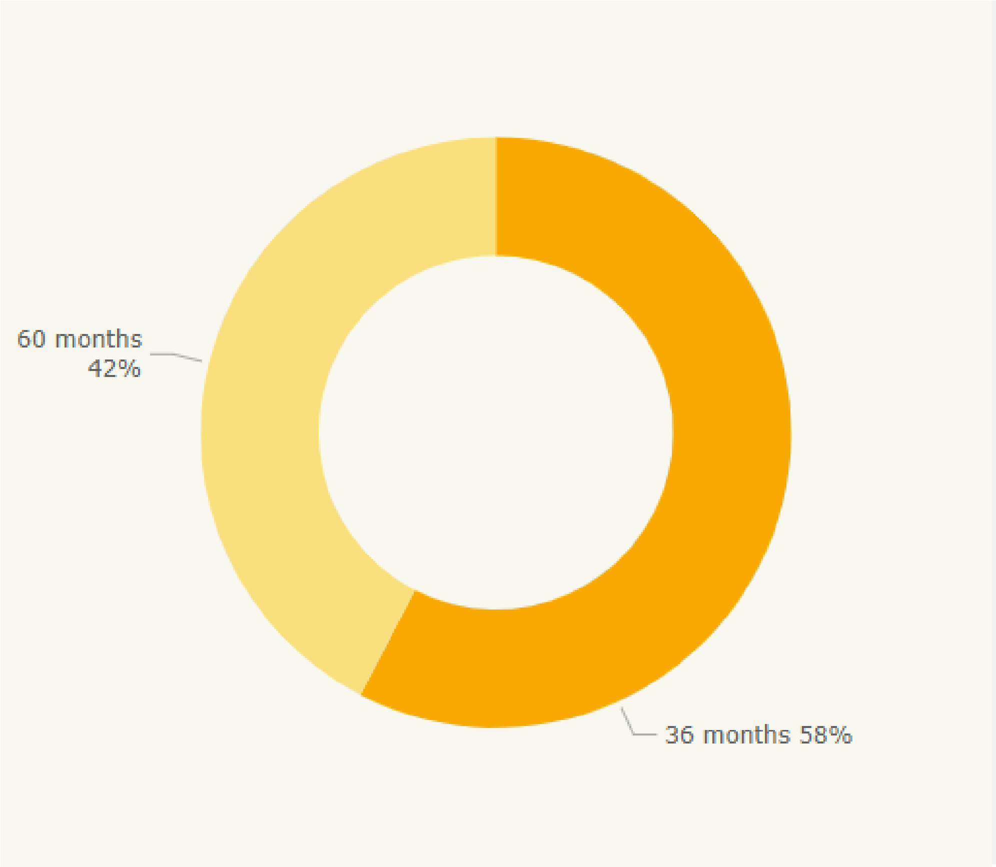
Potential categorical features: verification_status, term, type and grade.

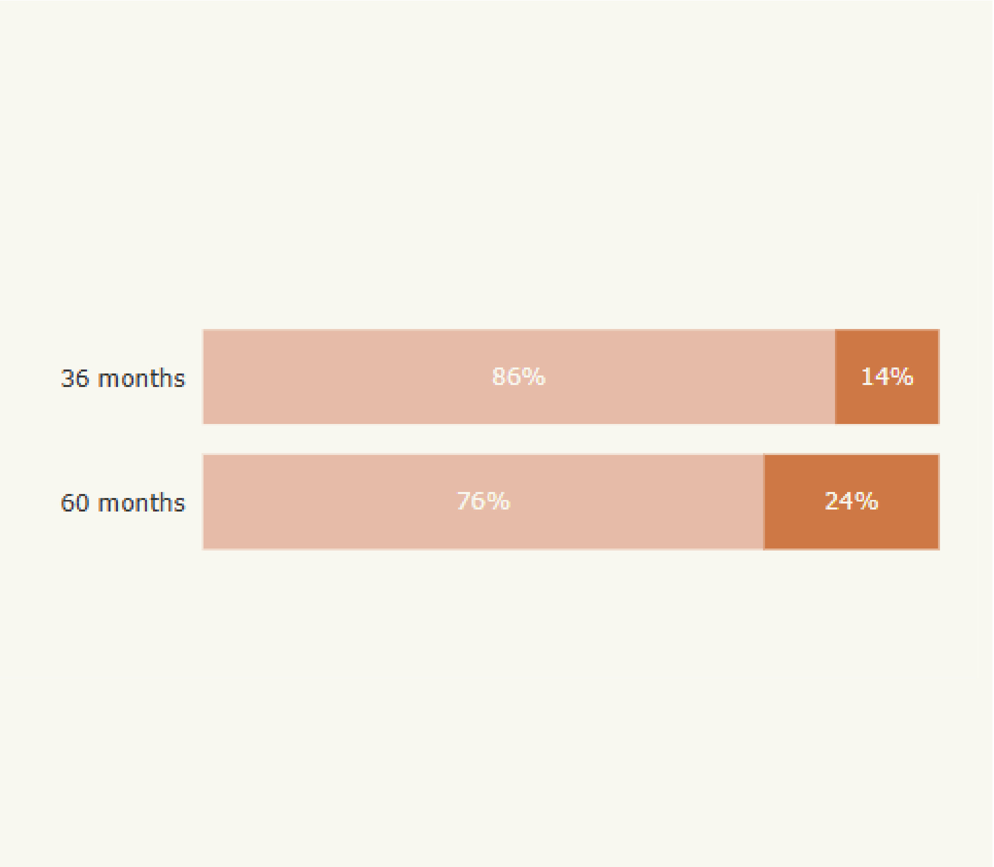
verification_status

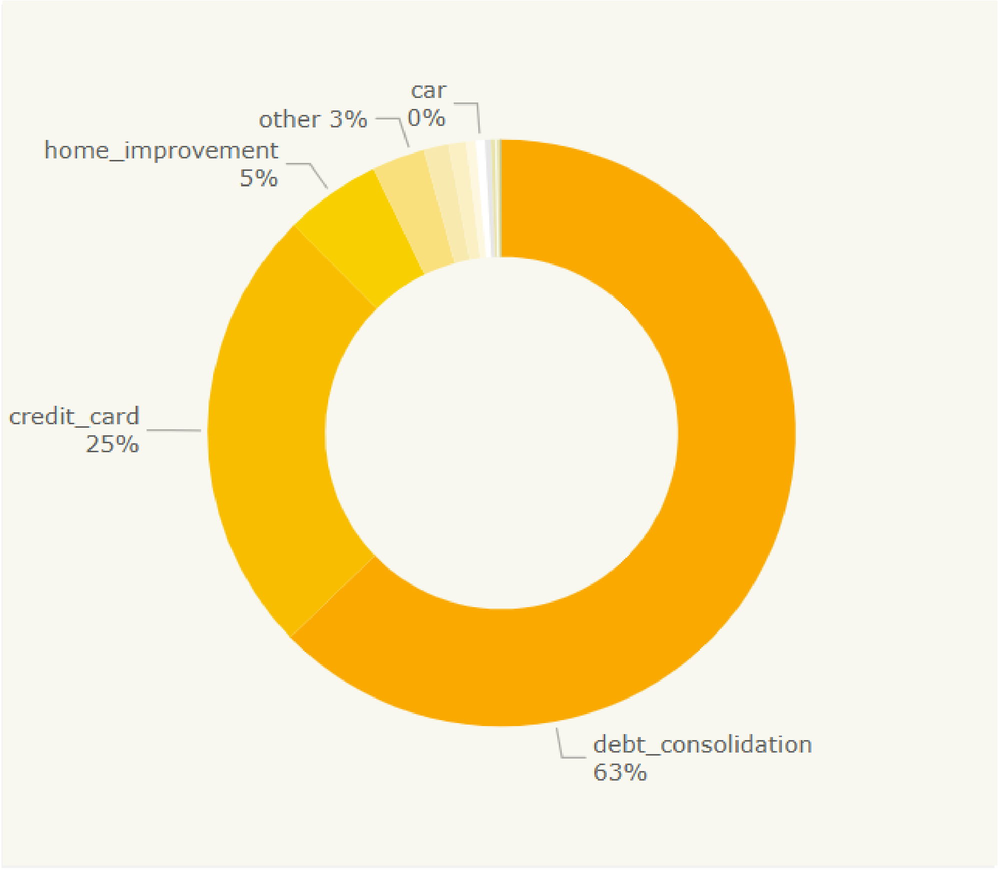


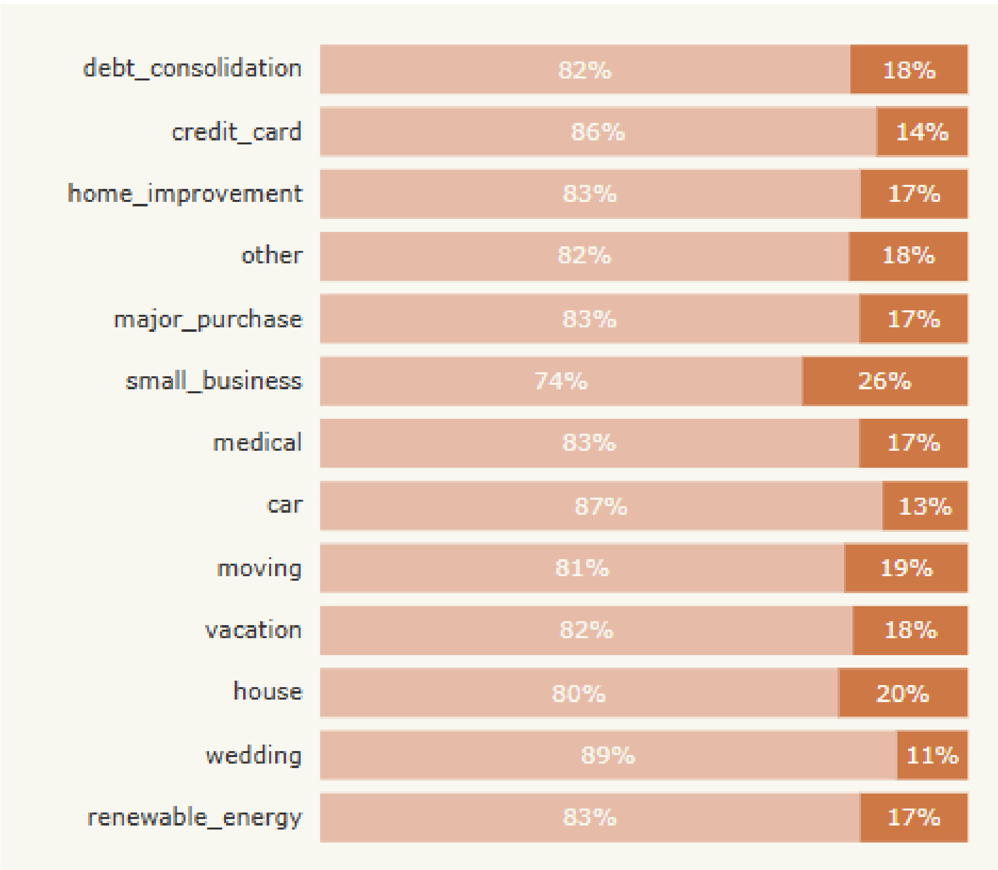


term

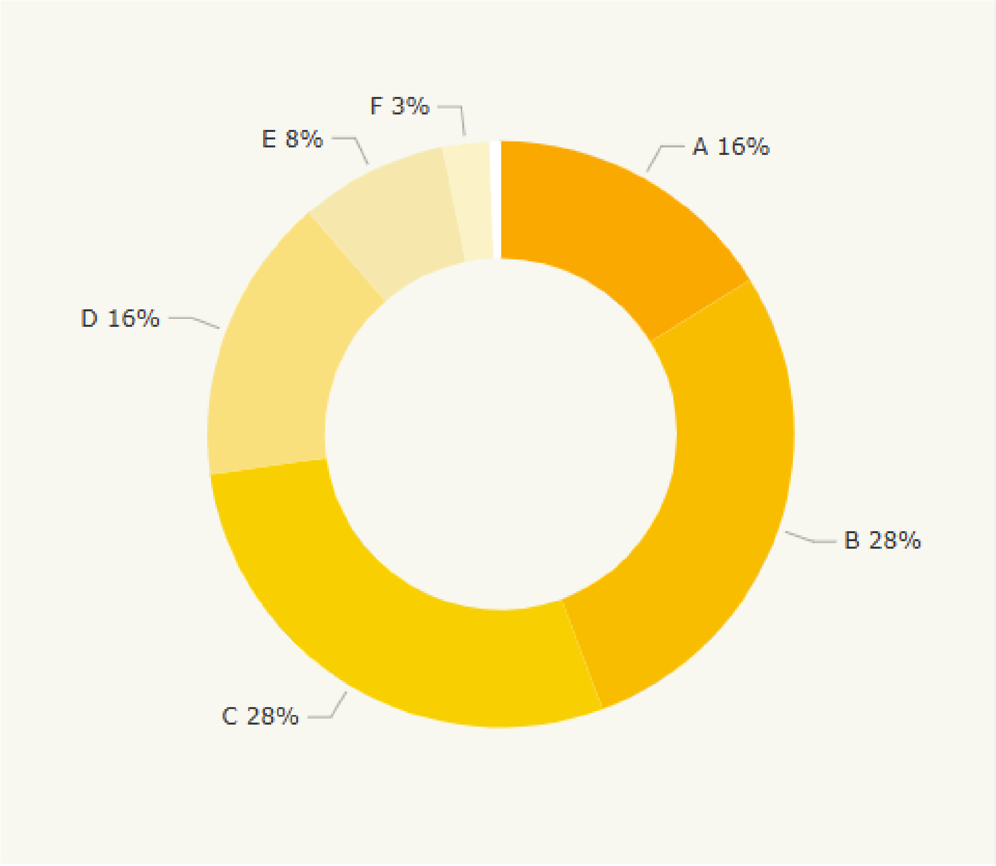


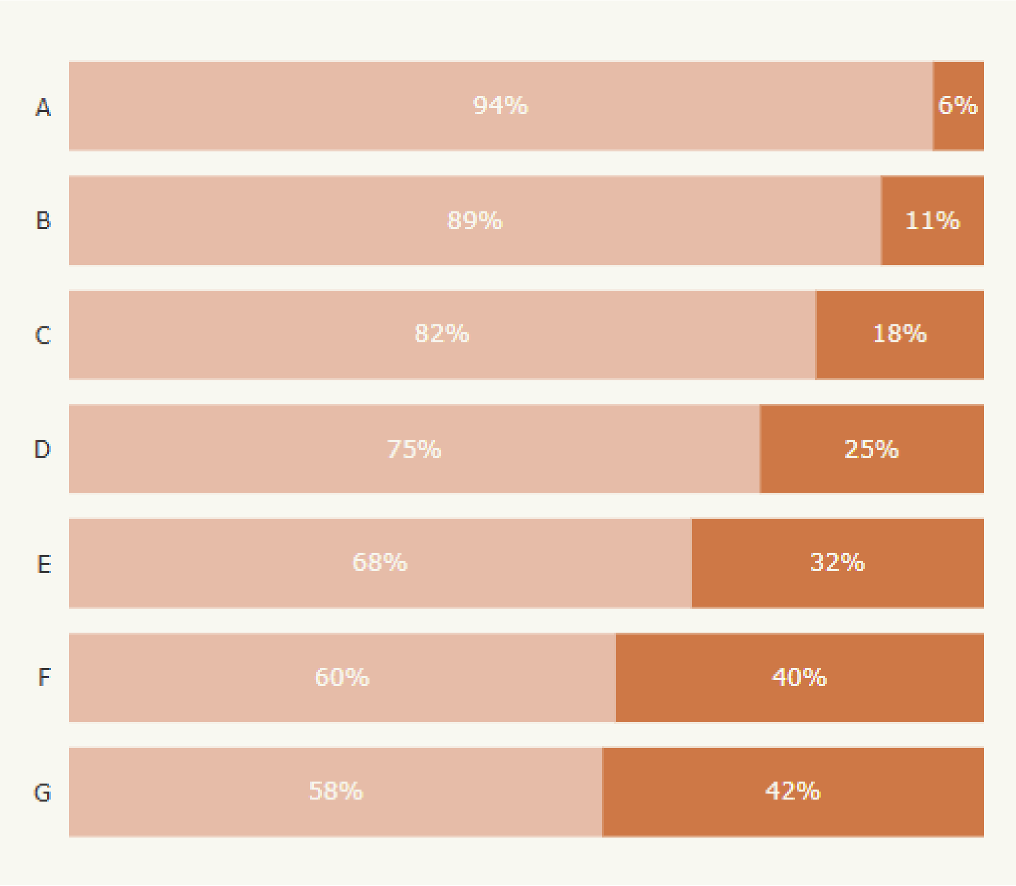






purpose





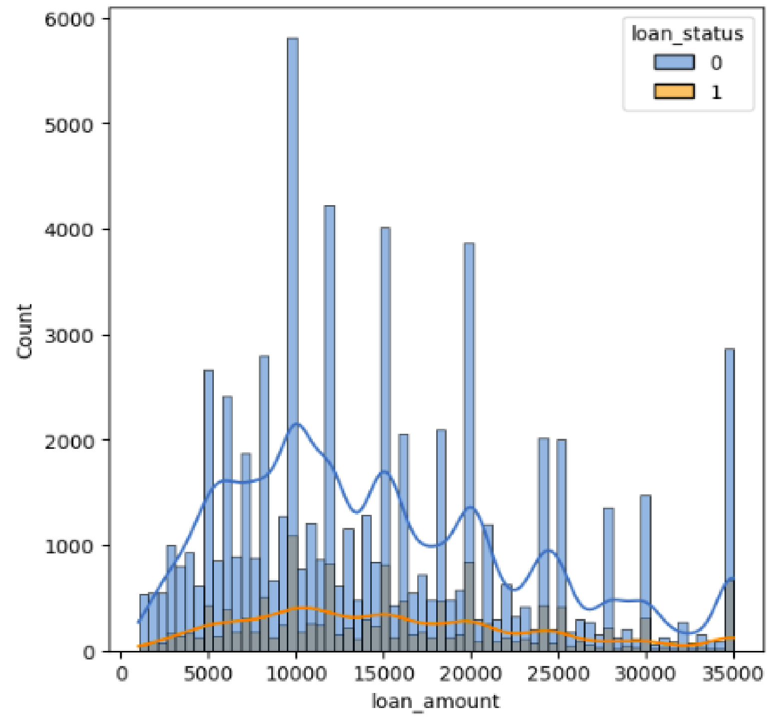
grade

Exploring Numerical Feature Distributions

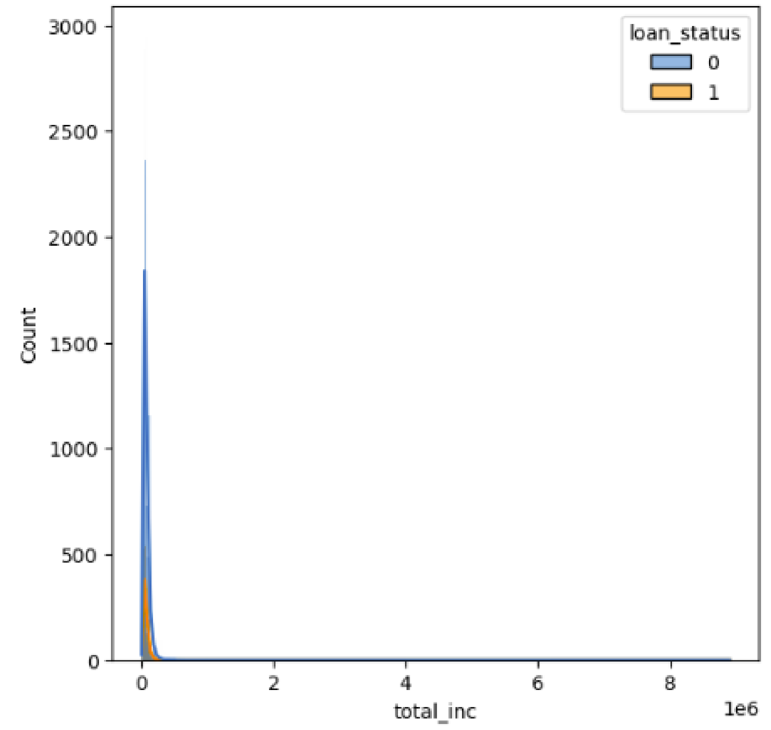
The distributions of **income** and **balance** variables are heavily skewed with many outliers. The class distributions overlap significantly, making them difficult to distinguish — these variables need to be transformed using StandardScaler.

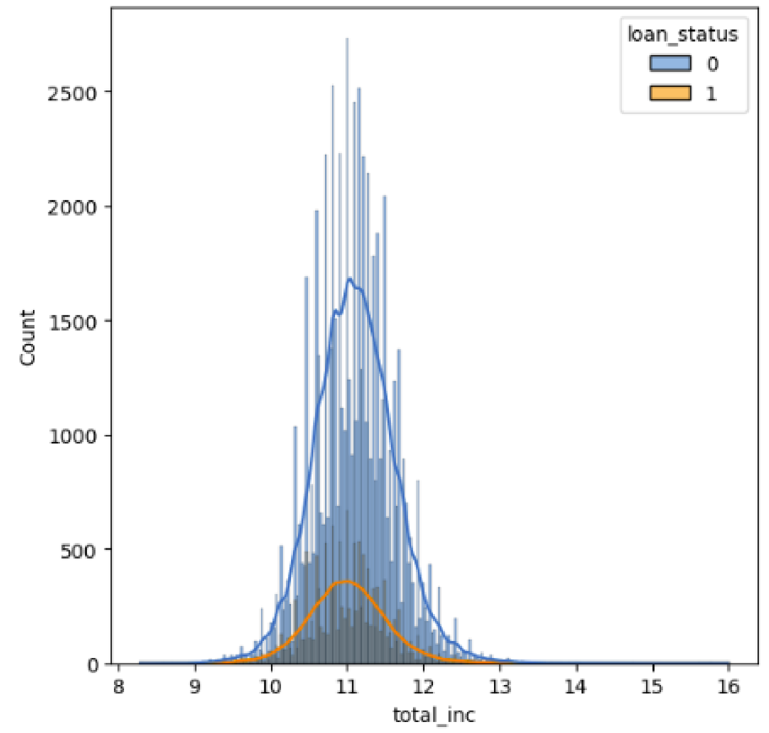
Among all numerical variables, only **int_rate** (interest rate) clearly shows a distinction between the two classes: Higher interest rates → higher-risk borrowers → more likely to have late payments.

loan_amount

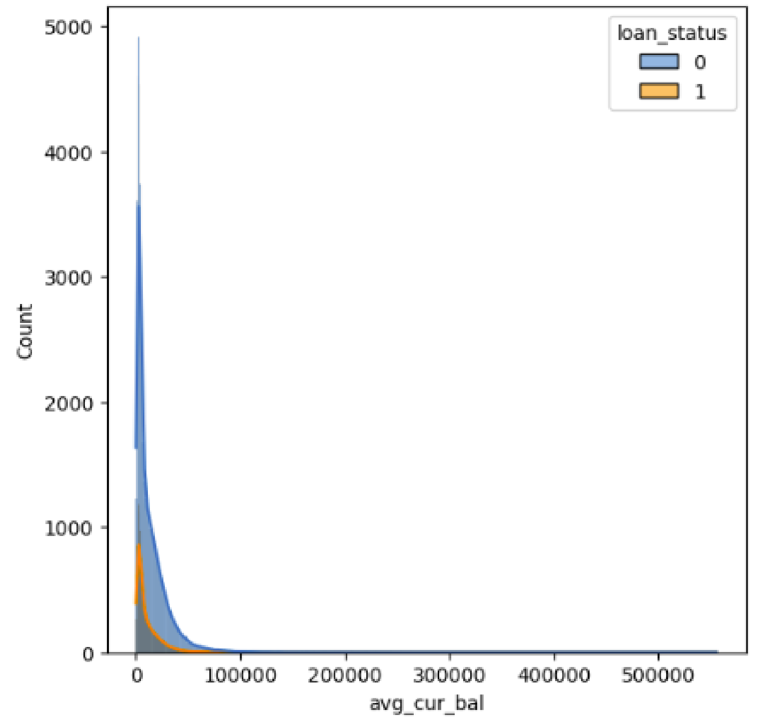


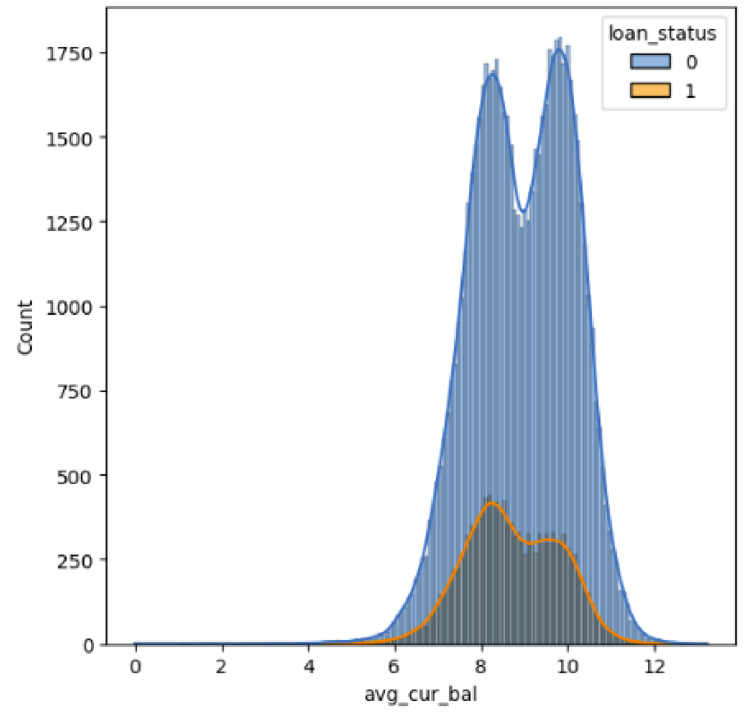
total_inc

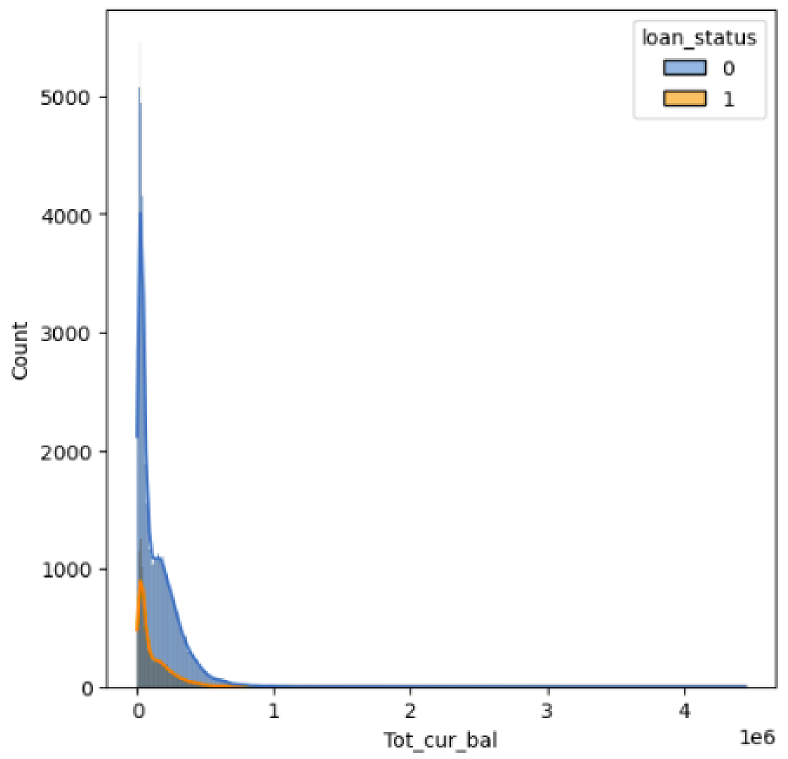


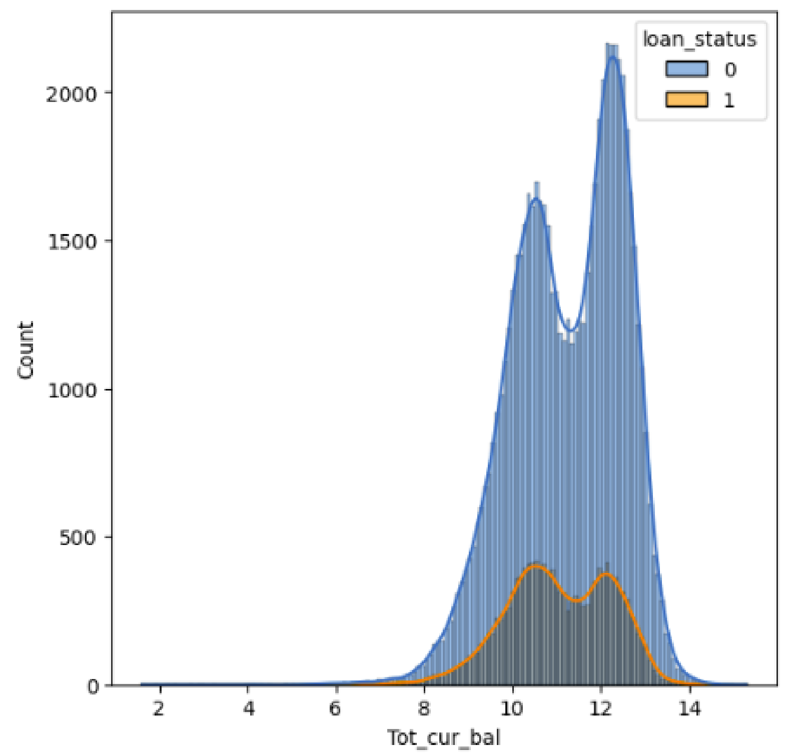


avg_cur_bal

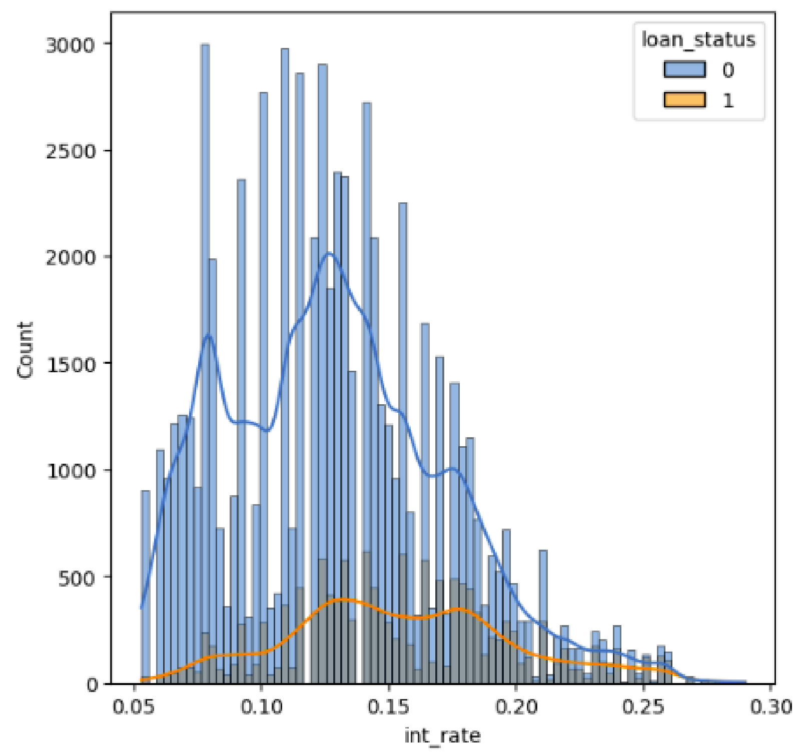


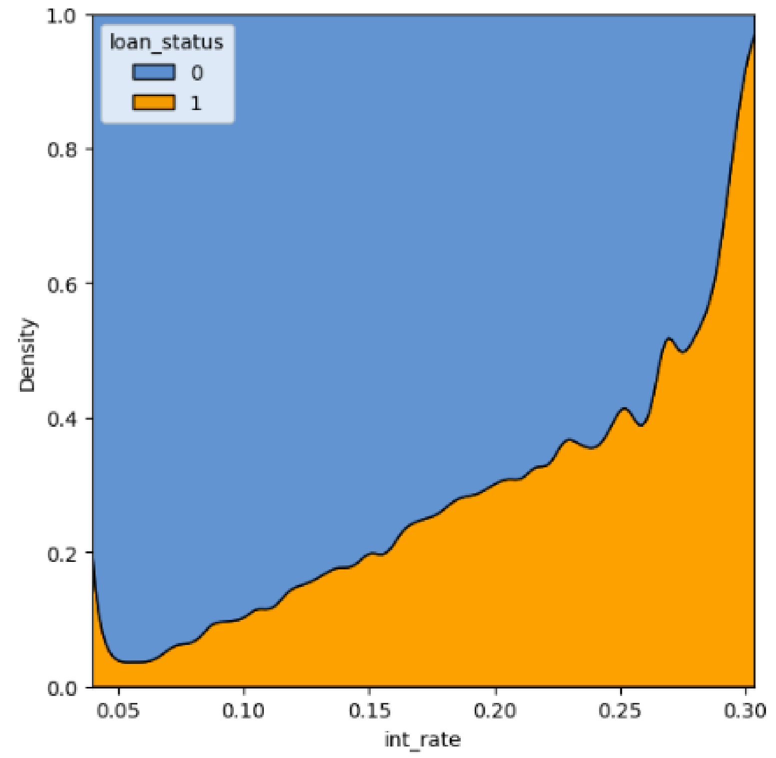




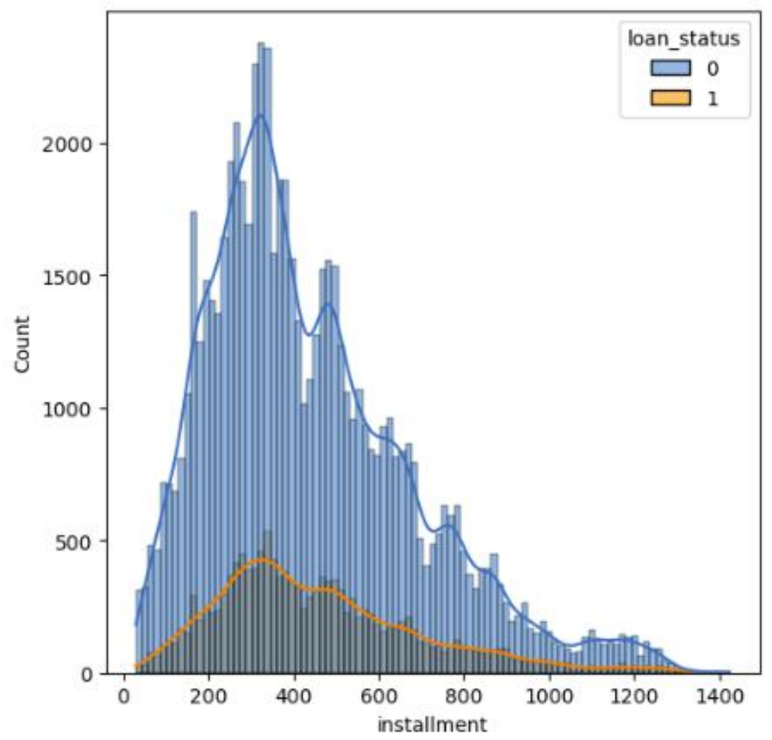


tot_cur_bal



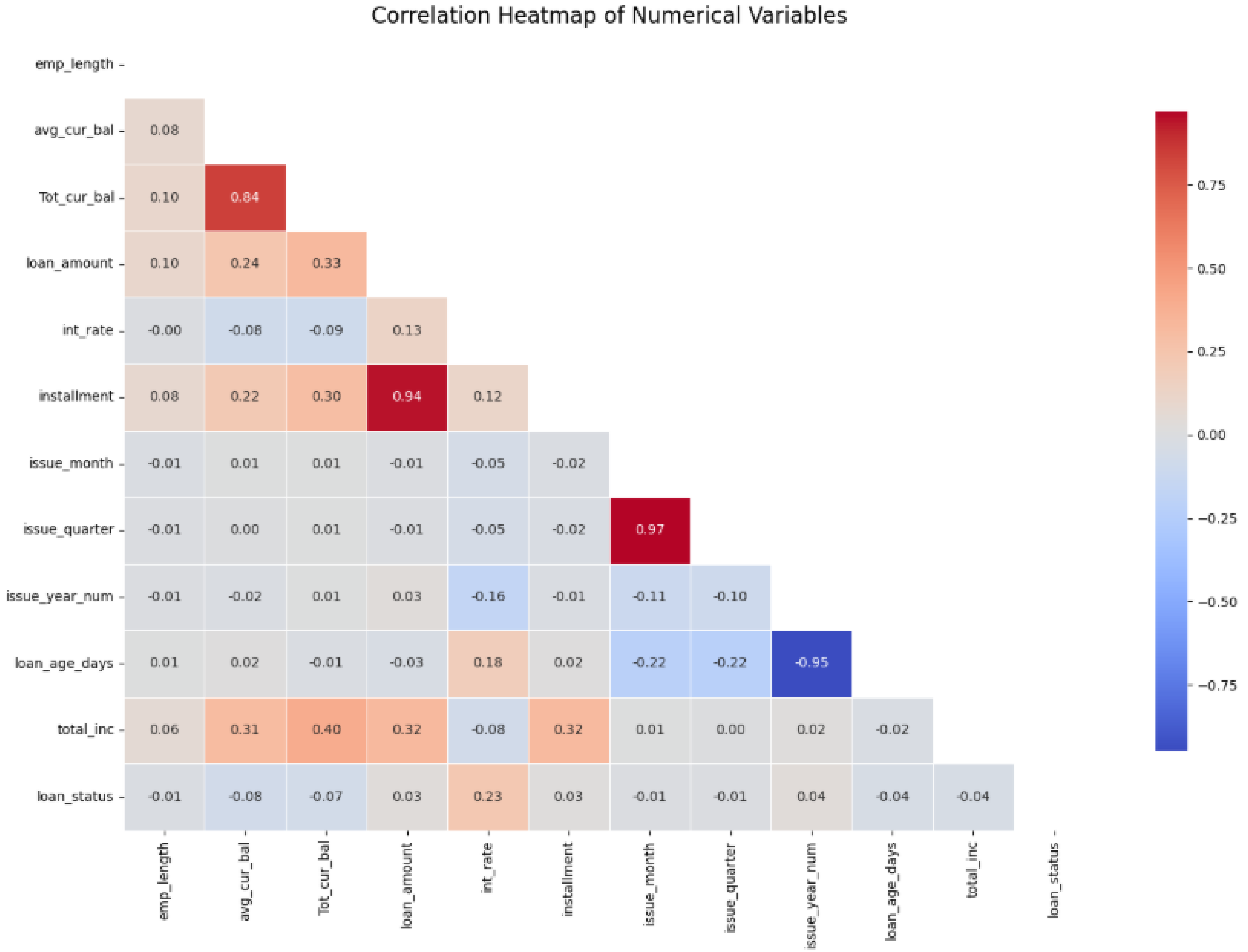


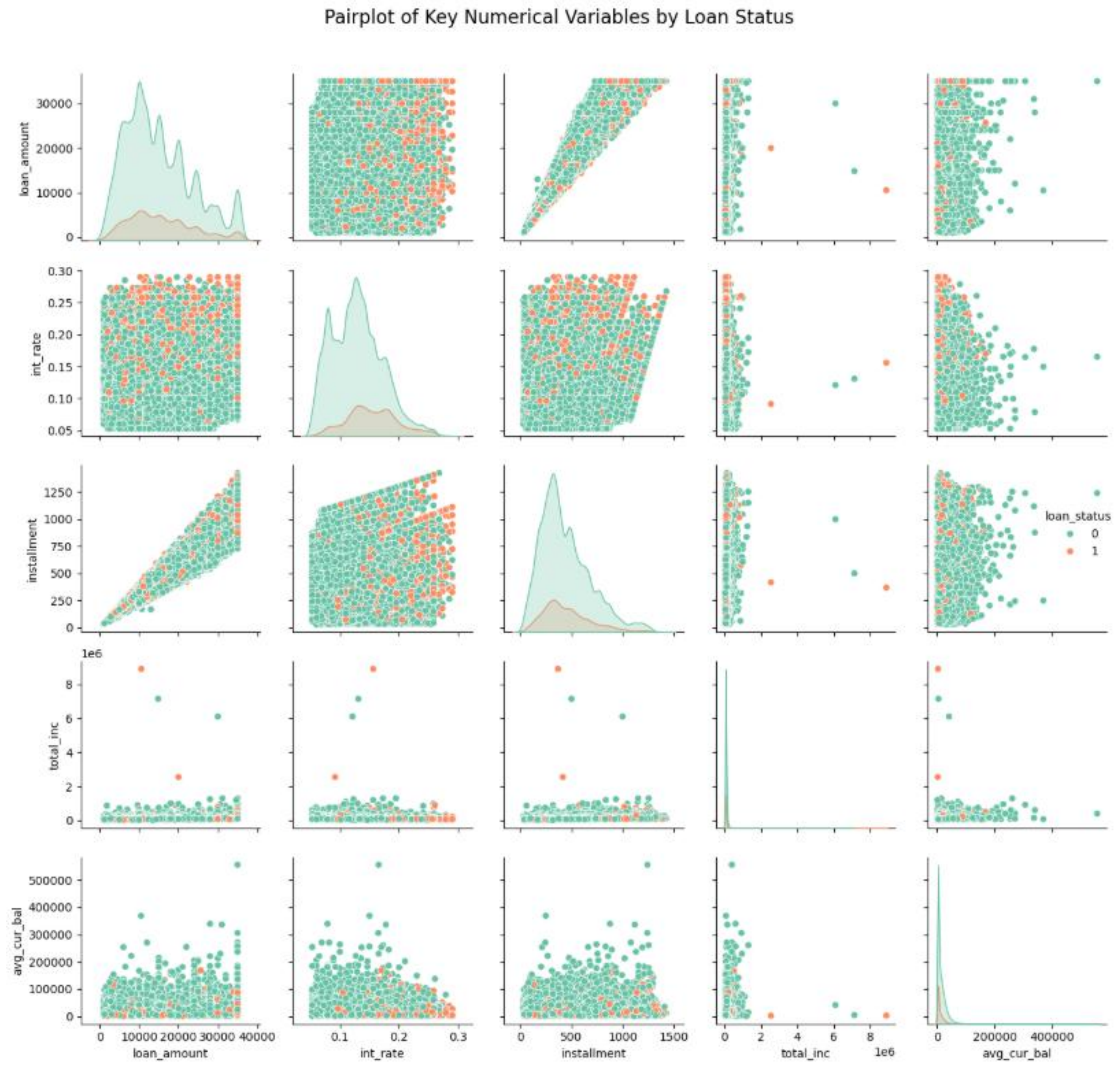
int_rate



installment

Multivariate Analysis





Conclusion:

✓ **'int_rate'** is the most individually predictive variable

✓ Relying solely on these raw features and their simple pairwise relationships will be insufficient to clearly distinguish between late payment and on time payment loans due to significant overlap.

New Features For Better Risk Prediction

✓ Raw features alone do not fully capture **behavioral signals** or **risk patterns** → created **13 new features** based on domain knowledge.

✓ These features significantly improved model performance: Feature importance plots showed **12/13** features in **Top 20 predictors.**

Type	Feature name	How to calculate
Numerical features	loan_age_days	Nb of days from loan issuance date to reporting date
	payment_to_income	installment / (total_inc/12)
	lti	loan_amount / total_inc
	interest_burden	int_rate * loan_amount / total_inc
	bal_to_loan	avg_cur_bal / loan_amount
	bal_to_income	avg_cur_bal / total_inc
	rate_premium	int_rate - avg int_rate by grade
	total_debt_ratio	Tot_cur_bal / total_inc
	relative_loan_size	loan_amount / avg loan_amount by grade
	monthly_interest	(int_rate / 100 / 12) * loan_amount
	interest_to_payment_ratio	monthly_interest / installment
	loan_to_balance_ratio	loan_amount / Tot_cur_bal
Categorical features	is_maturity	'Yes': loan is matured, 'No': loan is not matured

Robust Feature Selection & Outlier Handling

Problem:

Not all features contribute positively to the model, some add noise or introduce bias.

Resolution:

✓ Implements multiple methods to evaluate feature importance:

- *Univariate Selection (ANOVA F-test):* Identifies features with strong individual relationships to the target.
- *Mutual Information:* Captures non-linear relationships between features and target.
- *Recursive Feature Elimination:* Iteratively removes less important features.
- *Model-based Feature Importance:* Uses Random Forest to rank features by importance.

✓ Feature selection:

- Summarize the results and select features that are chosen by at least two of the above methods.

✓ Handle outliers:

- Use IQR method to handle outliers

--- COMPARING FEATURE SELECTION METHODS ---

	univariate	mutual_info	rfe	model_based	Total
bal_to_income	1	1	1	1	4
installment	1	1	1	1	4
total_inc	1	1	1	1	4
int_rate	1	1	1	1	4
monthly_interest	1	1	1	1	4
bal_to_loan	1	1	1	1	4
purpose_debt_consolidation	1	1	1	1	4
interest_burden	1	1	1	1	4
avg_cur_bal	1	1	1	1	4
payment_to_income	1	1	1	1	4
lti	1	1	1	1	4
grade	1	1	1	1	4
interest_to_payment_ratio	1	1	1	1	4
loan_amount	1	1	1	1	4
term_ 60 months	1	1	1	0	3
relative_loan_size	0	1	1	1	3
total_debt_ratio	1	0	1	1	3
verification_status_Source Verified	1	0	1	1	3
issue_month	1	0	1	1	3
term_ 36 months	1	1	1	0	3
emp_length	0	1	1	1	3
rate_premium	0	1	1	1	3
issue_quarter	1	0	1	1	3
job_level_Entry Level	0	1	1	1	3
issue_year_num	1	0	1	1	3
loan_to_balance_ratio	0	1	1	1	3
loan_age_days	1	0	1	1	3
Tot_cur_bal	1	0	1	1	3
verification_status_Verified	1	0	1	1	3
home_ownership_RENT	1	1	0	0	2
home_ownership_MORTGAGE	1	1	0	0	2
profession_Business, Finance & HR	1	1	0	0	2
subregion_Middle Atlantic	0	0	1	1	2

Data Split & Data Pipeline

Train Test Split:

✓ Time-based split:

- Train set: From 08/2012 to 06/2015
- Test set: From 07/2015 to 12/2015

✓ Rationale for Choice:

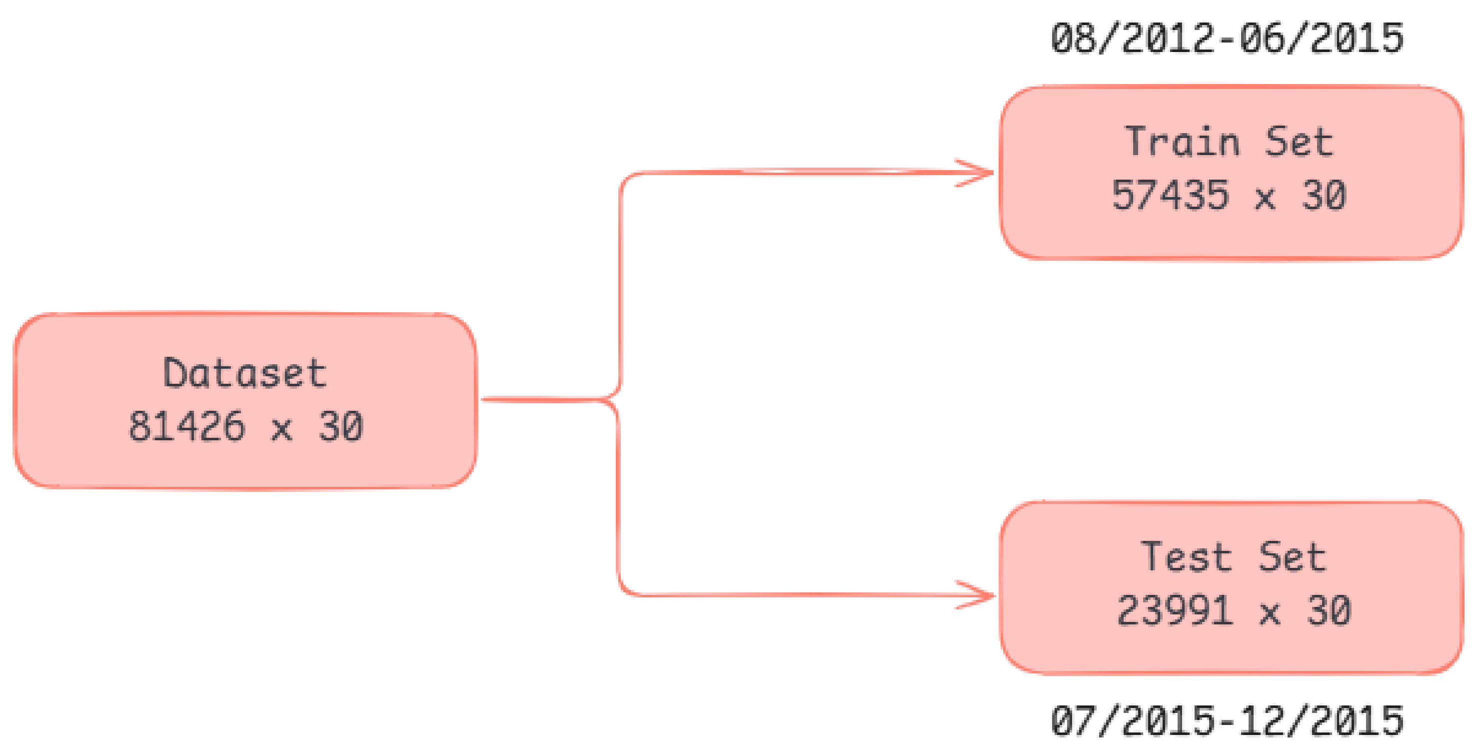
- Credit data is time-sensitive.
- A time-based split provides a more accurate assessment of how the model will perform on unseen data in future.

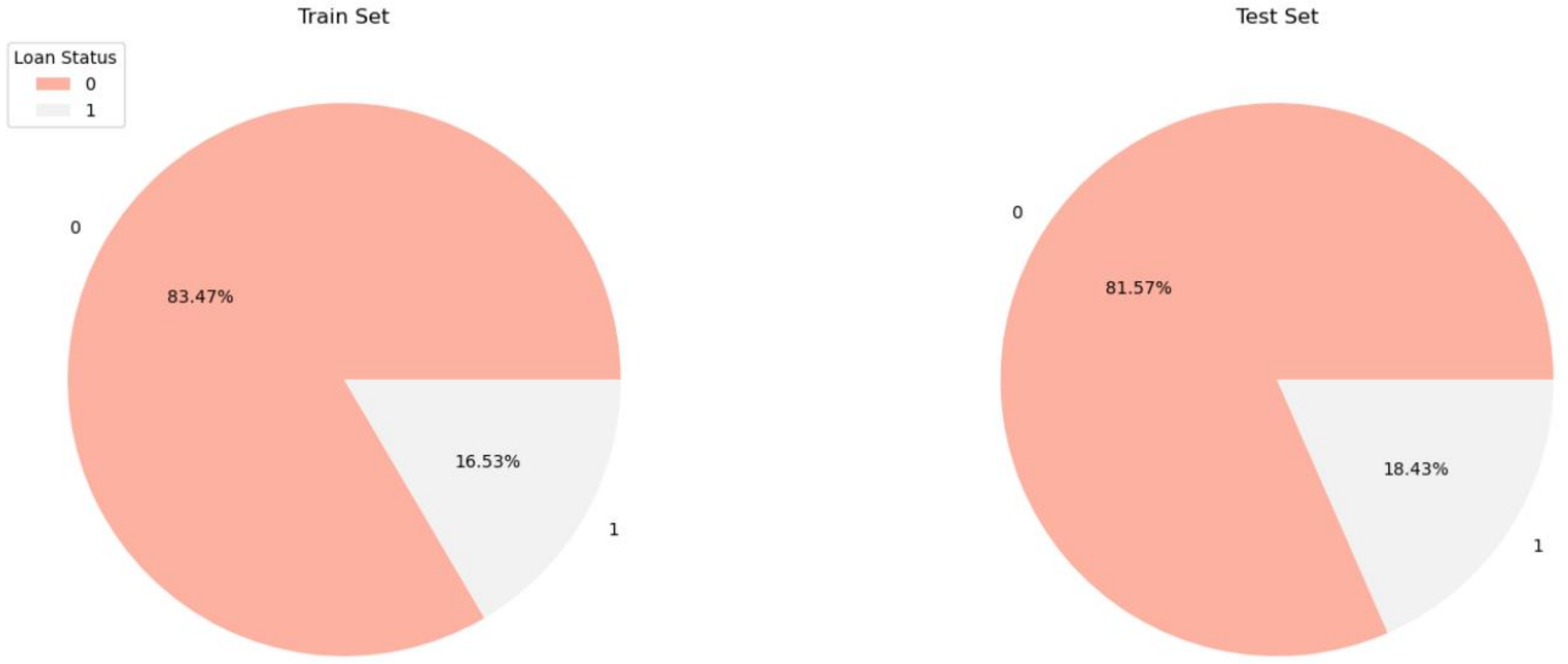
Data Pipeline:

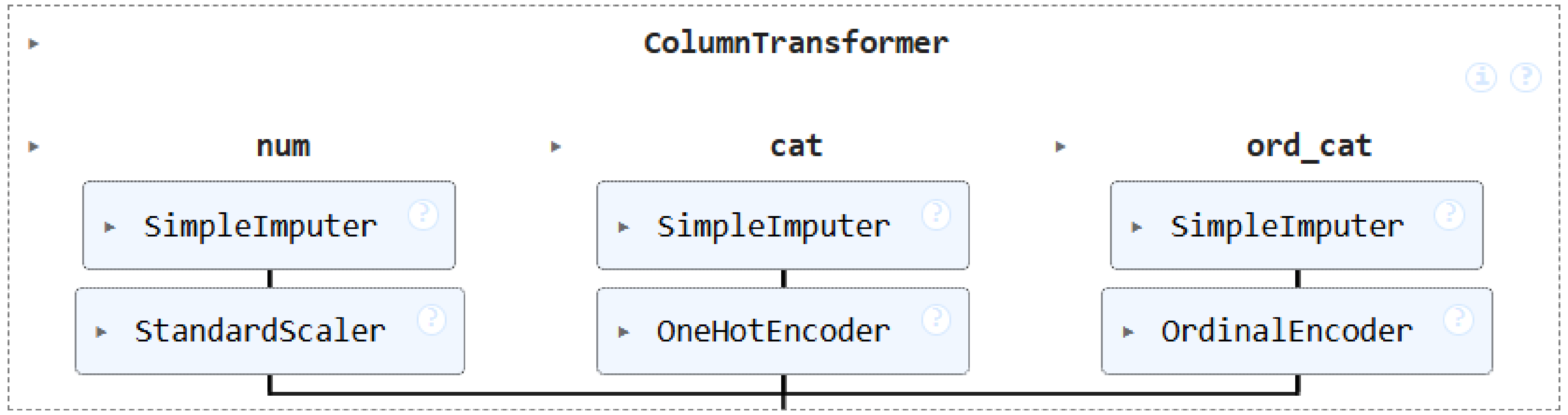
✓ Build an appropriate data pipeline for: numerical features, categorical features, and ordinal features.

✓ After the preprocessing process, the data increased from 30 columns to 67 columns due to the use of One-Hot Encoding on categorical variables.

✓ The original training data is almost dense, after preprocessing, the dataset has become highly sparse.







Bottom-up Approach: From Base Models to Final Ensemble

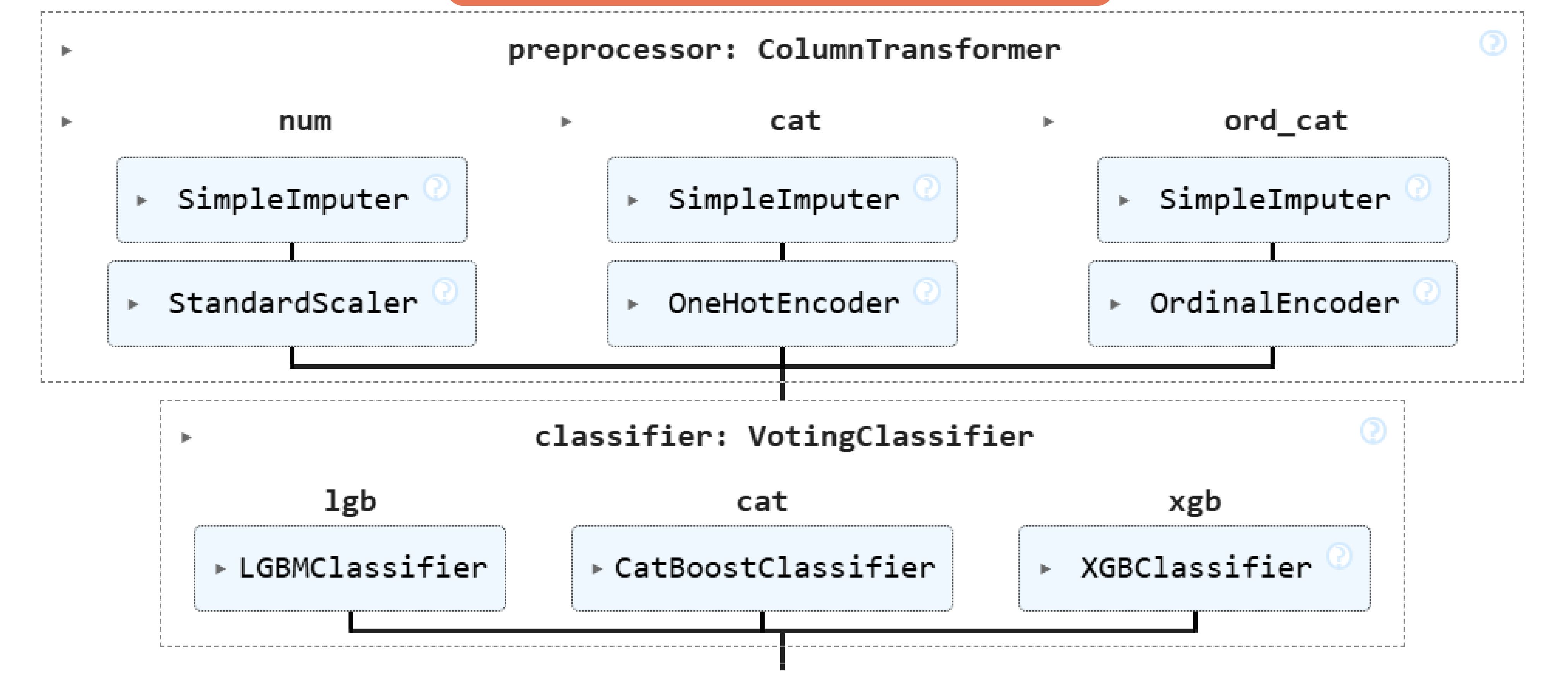
Followed a **2-phase** modeling strategy:

* **Phase 1:** Train and evaluate base models individually

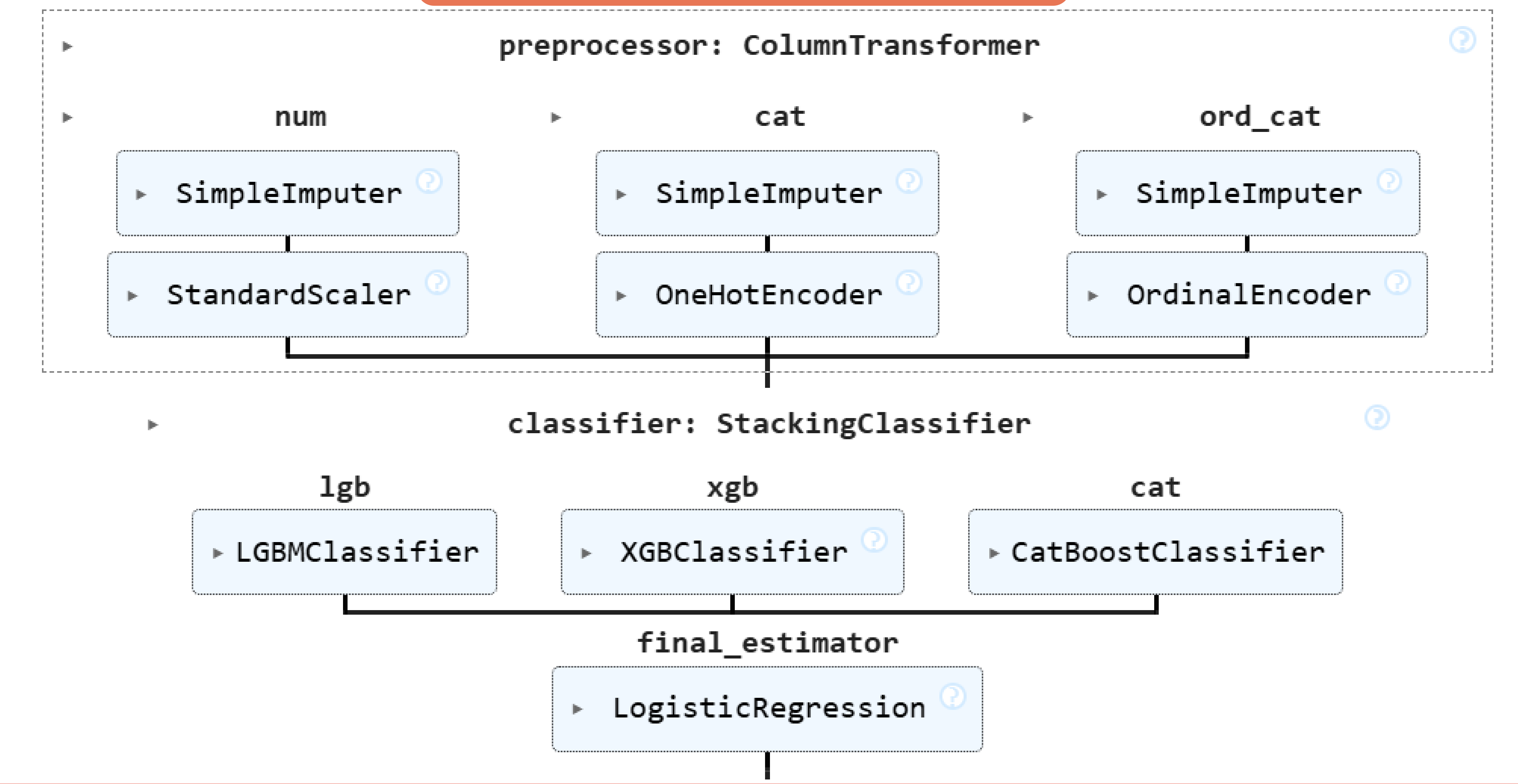
* **Phase 2:** Final Ensemble Model

Model	Precision	Recall	F1-score	Balanced-accuracy	PR-AUC	PSI
Logistic Regression	0.289458	0.659430	0.402318	0.646824	0.341633	0.0058
Balanced Random Forest	0.287975	0.625509	0.394382	0.638014	0.320649	0.3500
XGBoost	0.285672	0.654229	0.397691	0.642281	0.338969	0.0131
CatBoost	0.290726	0.657847	0.403244	0.647591	0.342556	0.0124
LightGBM	0.208091	0.947987	0.341271	0.566385	0.332106	0.0193

Weighted Voting Model



Stacked Ensemble Model



Model Performance: Which Model To Implement? Why?

Final Model Comparison

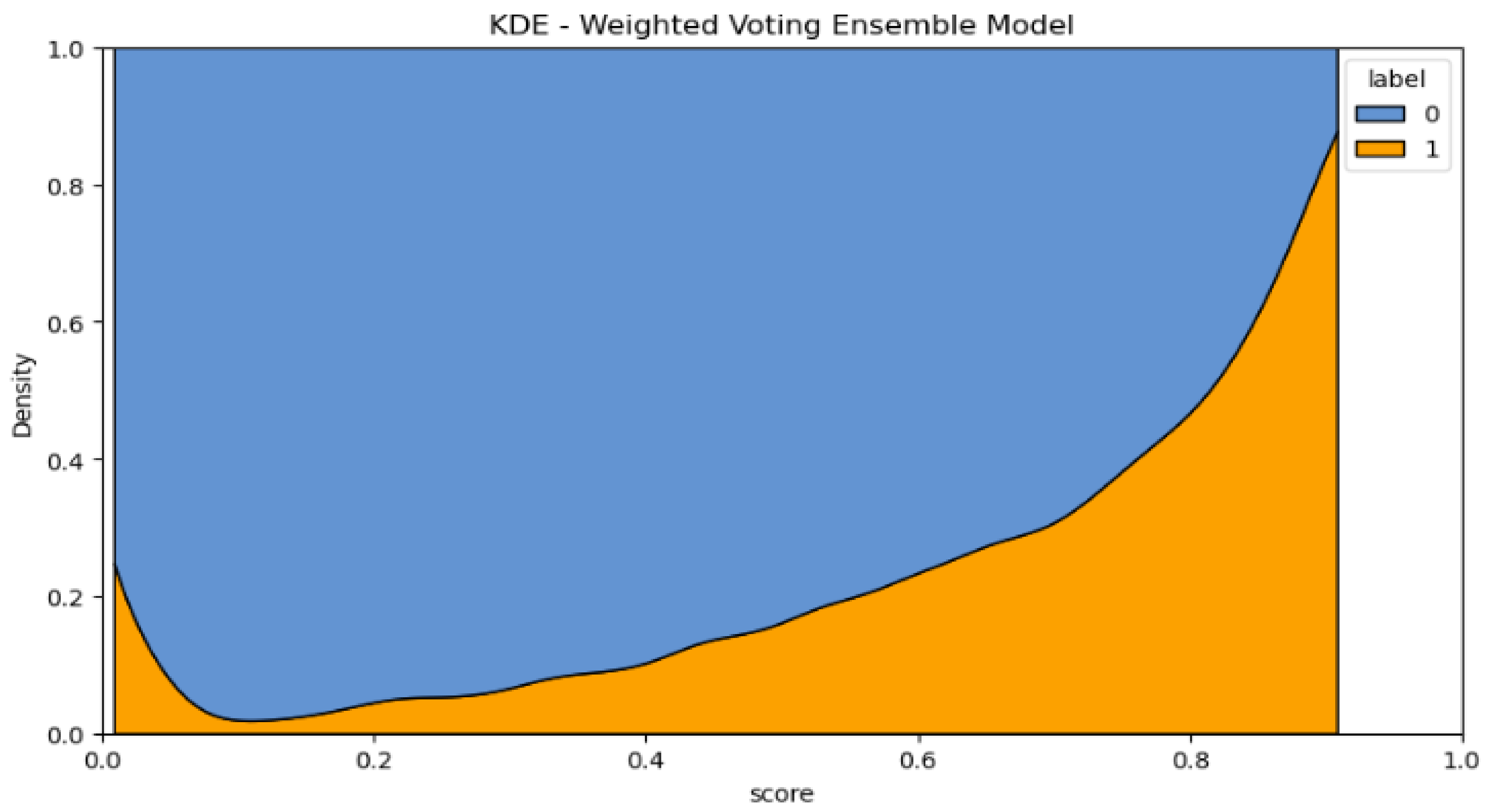
The best-performing models were selected to build the ensemble model, resulting in the summary table below:

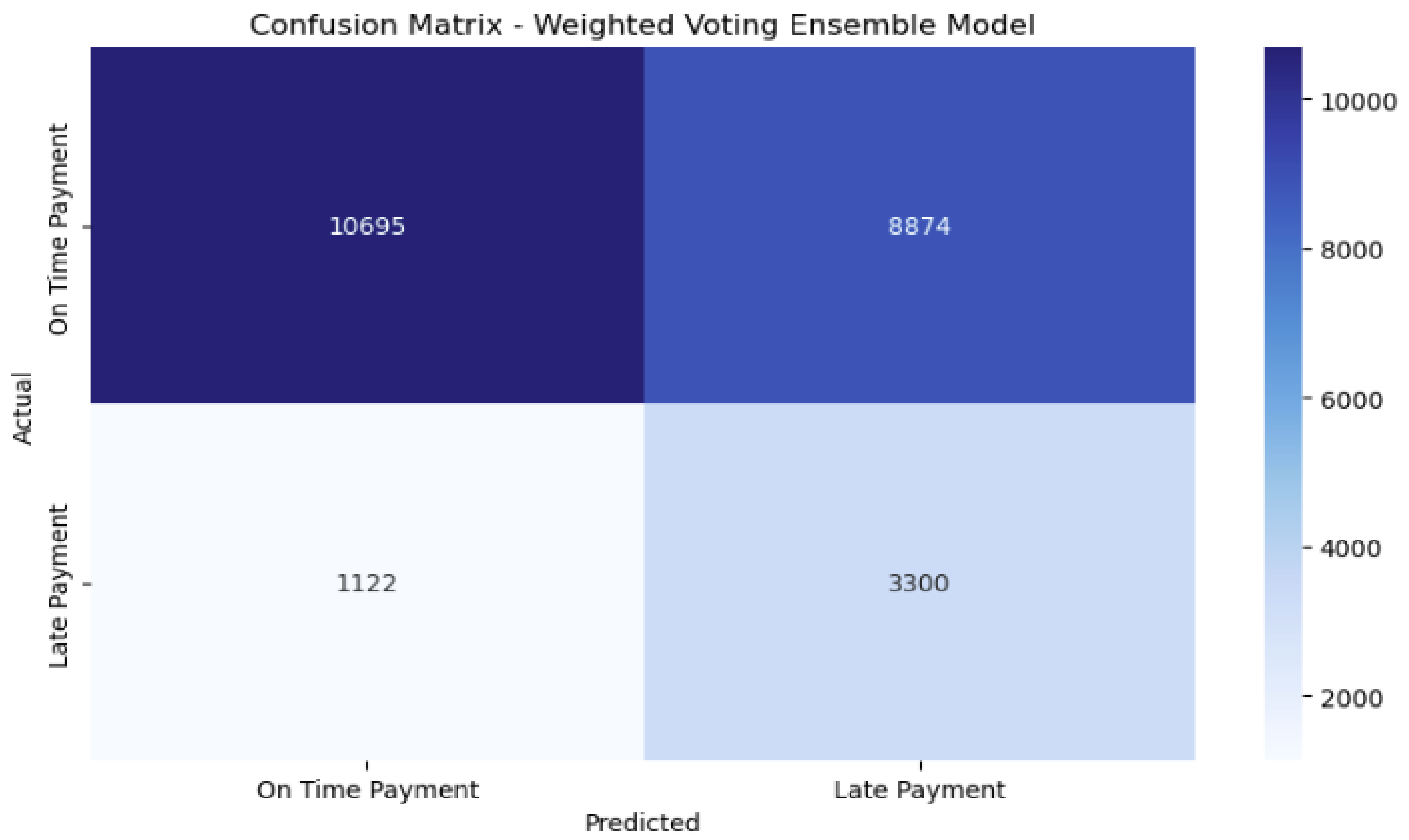
Model	Precision	Recall	F1-score	Balanced-accuracy	PR-AUC	PSI
Logistic Regression	0.289458	0.659430	0.402318	0.646824	0.341633	0.0058
Balanced Random Forest	0.287975	0.625509	0.394382	0.638014	0.320649	0.3500
XGBoost	0.285672	0.654229	0.397691	0.642281	0.338969	0.0131
CatBoost	0.290726	0.657847	0.403244	0.647591	0.342556	0.0124
LightGBM	0.208091	0.947987	0.341271	0.566385	0.332106	0.0193
Weighted Voting Ensemble Model	0.271069	0.746269	0.397686	0.646398	0.345080	0.0109
Stacked Ensemble Model	0.273269	0.735640	0.398505	0.646782	0.344765	0.0135

- ✓ **Key Metrics:** Recall and F1-score
- ✓ **Selected Model:** Weighted Voting Ensemble Model – Combining LightGBM, XGBoost, and CatBoost with voting weights in the order of [1, 4, 3]
- ✓ **Rationale for Choice:**
  - Highest PR-AUC (0.345080)
  - 2nd PSI (0.0109)
  - 2nd Recall (0.746269)
  - Competitive F1-score and balanced accuracy.

Weighted Voting Ensemble Model: Good But Not Good Enough

Model Performance





- ✓ There is a significant overlap between the blue and orange distributions in the middle score ranges leading to misclassifications.
- ✓ The model successfully identifies about 74.63% of the actual late payments but when the model predicts a late payment, it's correct about 27.11% of the time.

Can we accept a cost of approximately 3 false alarms for every 1 correctly identified 'true late payment'?

Hyperparameter Tuning – Small F1 Gain, But Big Recall Drop

Approach:

✓ Used Optuna to tune key parameters of LightGBM, CatBoost, and XGBoost in ensemble model.

Objective:

✓ Improve F1 score

Result:

✓ F1 Score: ↑ 0.49%

✓ Recall: ↓ 8.65%

Decision:

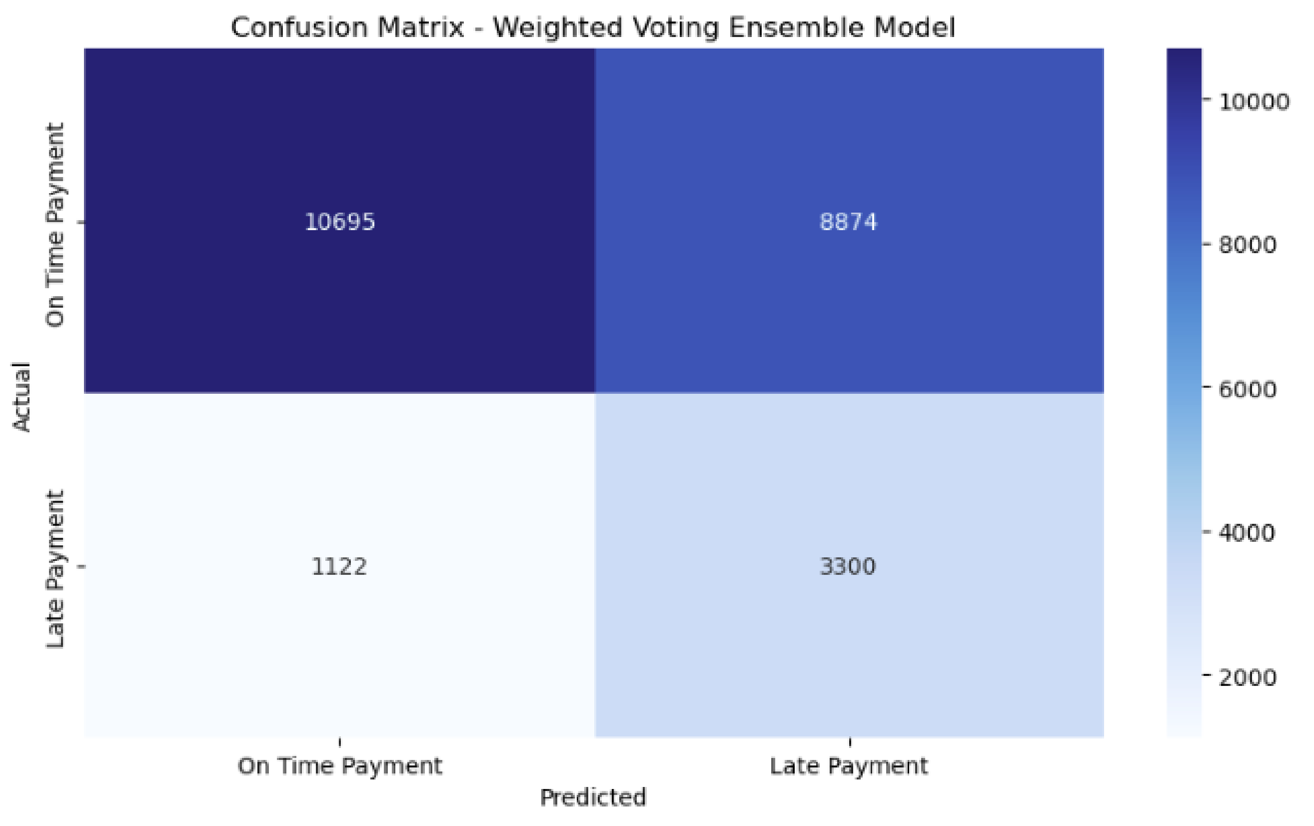
Despite marginal F1 gain, the drop in recall means more risky loans could be missed, which conflicts with business risk management goals

➢ **Used the original model.**

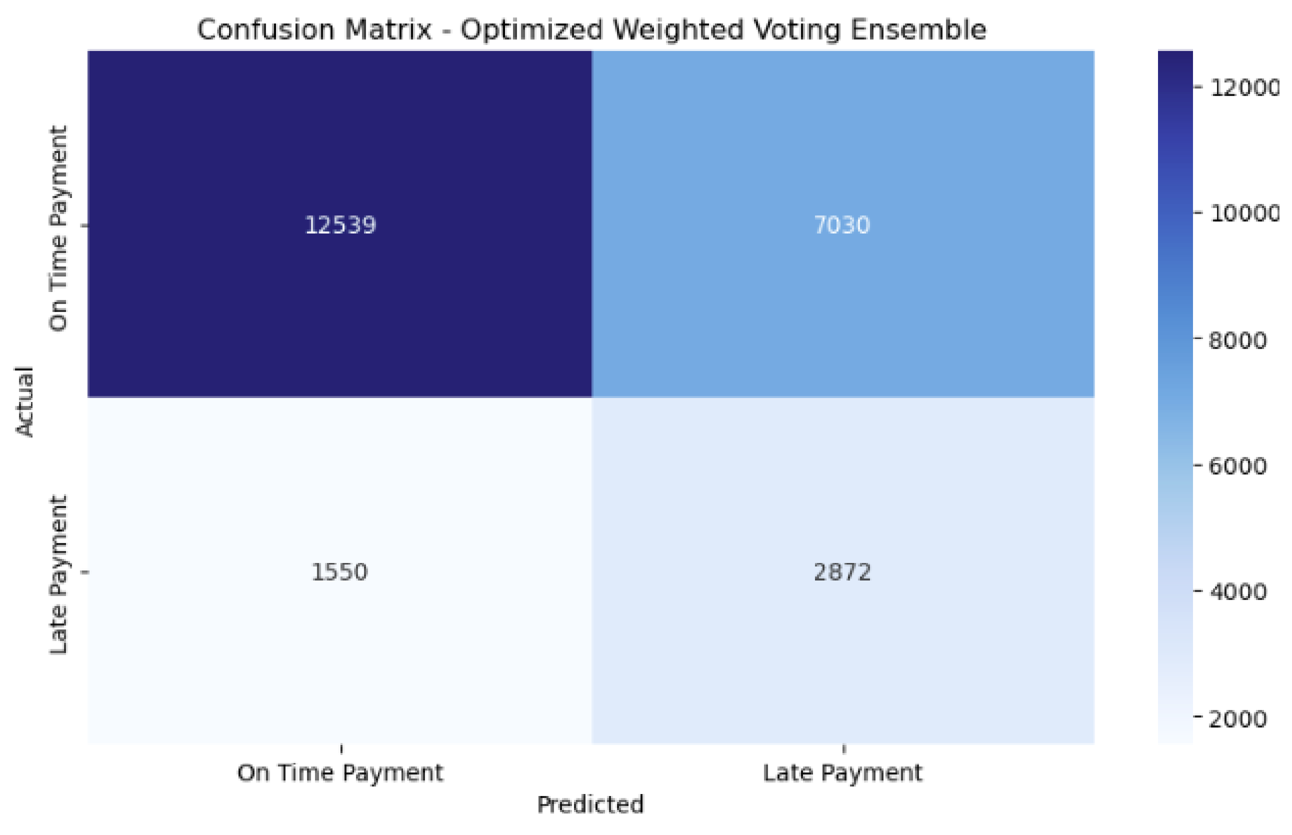
Evaluate performance after tuning

Model	Original Model	Optimized Model
Precision	0.271069	0.290420 ↑
Recall	0.746269	0.649480 ↓
F1-score	0.397686	0.399061 ↑
Balanced-accuracy	0.646398	0.646210 ↓
PR-AUC	0.345080	0.340389 ↓

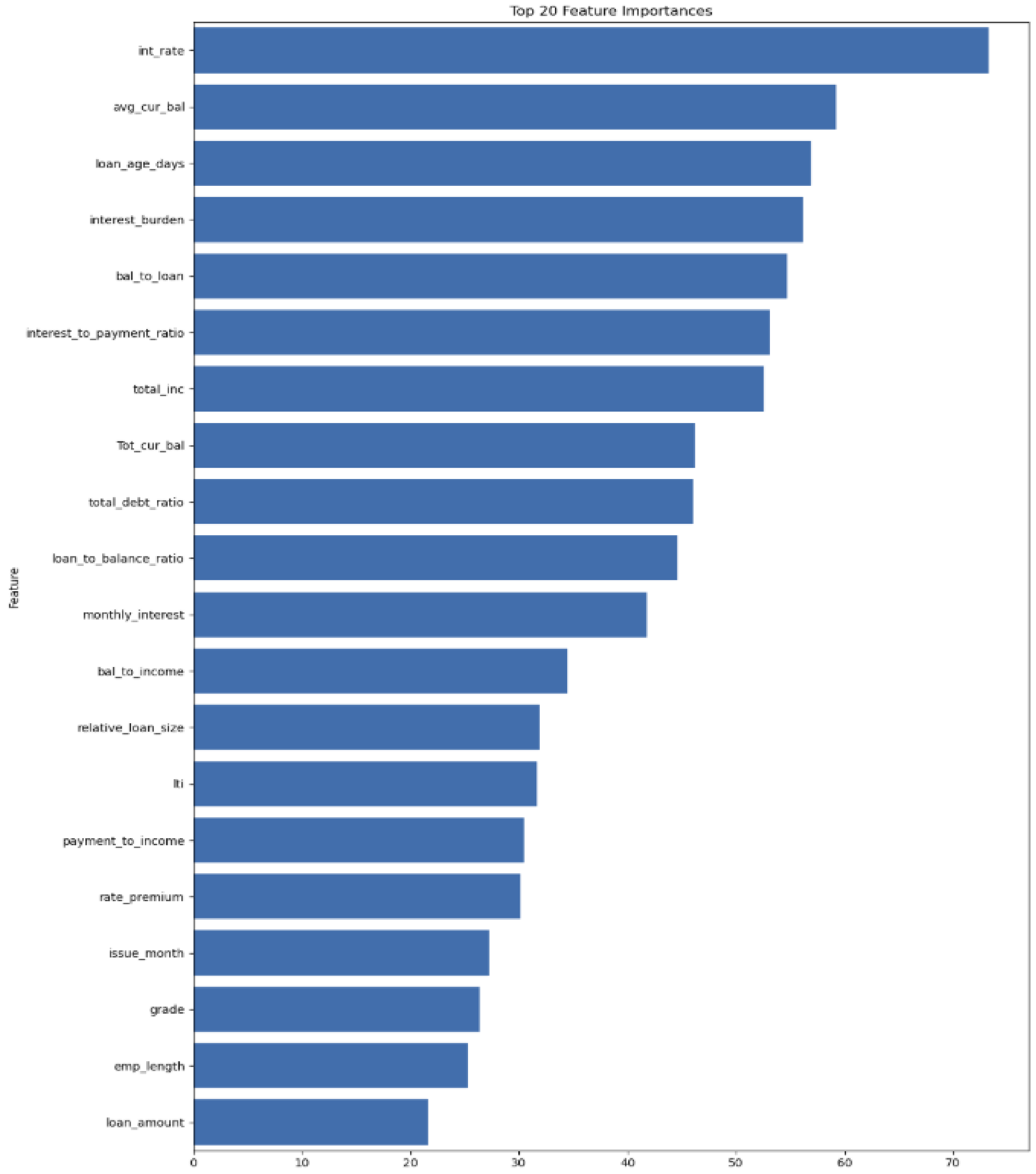
Before



After



Interest Rate & Finance-Related Factors Have The Most Influence



Key Drivers of Late Payment Prediction

- ✓ **Interest Rate** (int_rate) is the most critical factor, indicating a strong correlation between higher rates and increased late payment risk.
- ✓ **Average Current Balance** (avg_cur_bal) also significantly influences the prediction, suggesting that the current financial burden on the borrower is a key indicator.
- ✓ loan_age_days, interest_burden, and various **financial ratios** also have a significant impact.
- ✓ **Demographic factors** such as occupation, region, etc., have less impact compared to the variables mentioned above.

➢ It is necessary to **focus on financial factors** when evaluating new loans and develop more targeted proactive intervention strategies, thereby directly impacting loss reduction and improving portfolio health.

From Prediction to Action: Operationalizing Model Results

Which department will the model results be sent to?

✓ Credit Risk Management Department.

✓ Collections Department.

✓ Loan Underwriting Department.

✓ Customer Relationship Management (CRM) Team.

How to operate the results from the model?

✓ **Scenario:** A new loan application is processed, or an existing loan is periodically reviewed.

✓ **Input:** Applicant/Loan data is fed into the model.

✓ **Output:** The model classifies each customer: 'At Risk of Late Payment' - Yes or No, along with a probability score to serve as a risk assessment scale.

✓ **Actionable Steps based on Risk Level:**

- 'At Risk - No': Standard monitoring. No immediate proactive intervention required based on this model.
- 'At Risk - Yes':
  - Segment further based on the predicted probabilities (High Risk, Medium Risk, Low Risk)
  - Action: Proactive outreach by the Collections/Customer Care team before the payment due date.
    - Friendly payment reminders via SMS/Email.
    - Offer to set up automatic payments.
    - For very high-risk 'Yes' cases: A direct call to discuss upcoming payment.

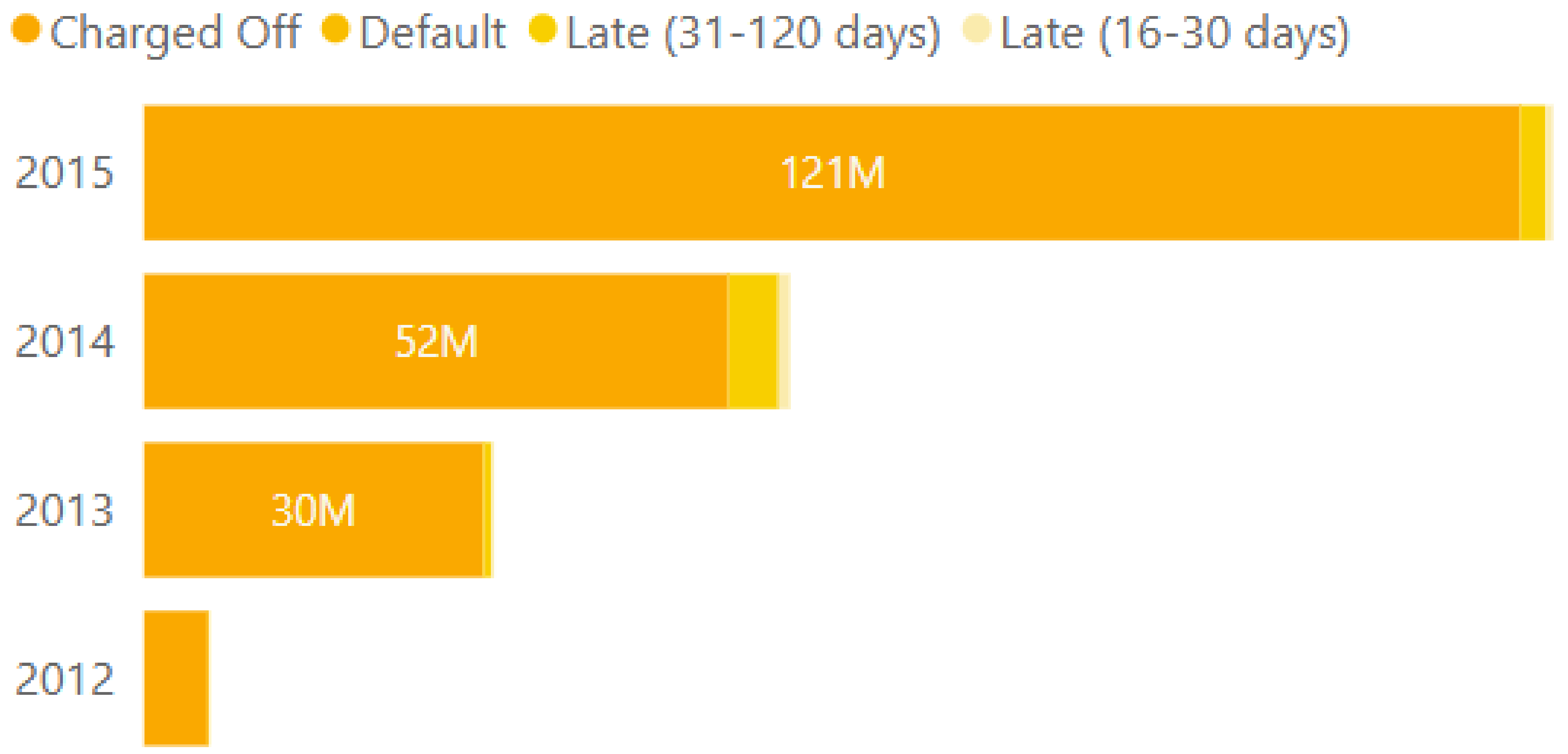
Integration:

✓ Monthly list of 'At Risk - Yes' customers sent to CRM for automated outreach sequences

✓ Flag in the loan servicing system for account managers

Anticipated Business Impact: Reducing Late Payments

Late payment amount by year



Problem:

In recent years, the number of loans issued by the organization has increased rapidly, reflecting strong growth in the organization's scale. However, the organization has also faced increasing losses from late payment loans. Specifically, in 2015, the organization suffered a loss of $121 million from charged-off loans.

Complication:

Quantifying the direct financial impact of a predictive model can be challenging without concrete examples.

Action	Business Impact
Proactively identifying and intervening with 20% of high-risk loans	Assumption: Reduce the overall default rate by 10%. Estimated Impact: Reduce the defaults by $12.1 million.
Identifying at-risk loans earlier and implement more efficient intervention strategies	Reduce the rate of loans turning into defaults, potential reduction in provisions for bad debt
Proactive and supportive communication with customers who might be facing temporary difficulties	Reduce churn, improve goodwill and foster stronger relationships
Credit risk and collections teams focus their efforts on accounts that are truly at risk.	Improving efficiency and reducing operational overhead.
Proactive manage loan portfolio	Enhancing the bank's reputation and attracting more creditworthy borrowers.

Future Opportunities: Expanding Predictive Capabilities

Limitations:

- ✓ Low precision leads to a significant number of false positive results.

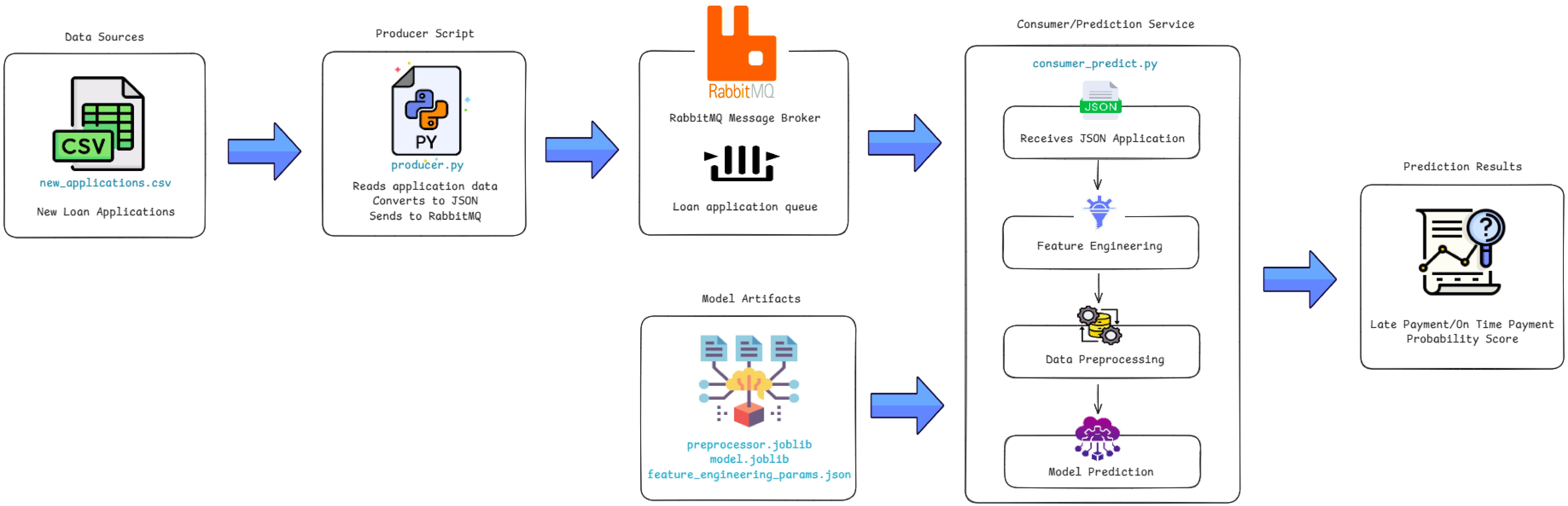
Other Methods or Problems that can be Expanded:

- ✓ Model Refinement: Continue to refine features that best separate the 'At Risk - Yes' and 'No' classes.
- ✓ Predicting Likelihood of Multiple Late Payments: Move from a single late payment prediction to identifying chronic late payers.

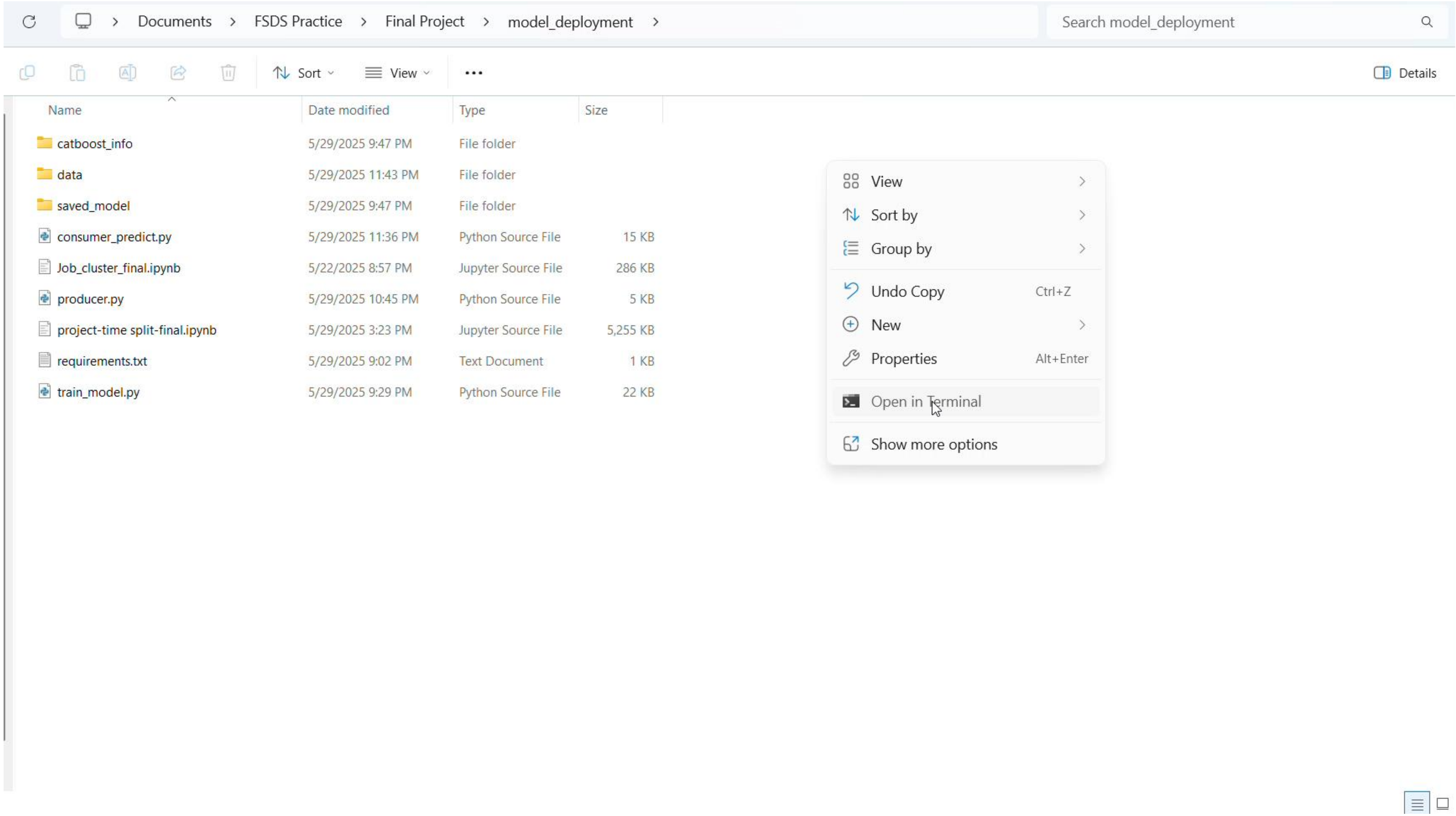
Business metrics that can be related:

- ✓ Net Charge-Off Rate
- ✓ Provision Rate.
- ✓ Cost of Collections per Account
- ✓ Customer Churn Rate
- ✓ Operational Efficiency Metrics, such as manual review hours for credit analysts or outbound collection calls…

Workflow Diagram Description



Model Deployment – Demo Video



APPENDIX

Data Selection Justification

Why Only Used 2012–2015 Data for Model

Situation

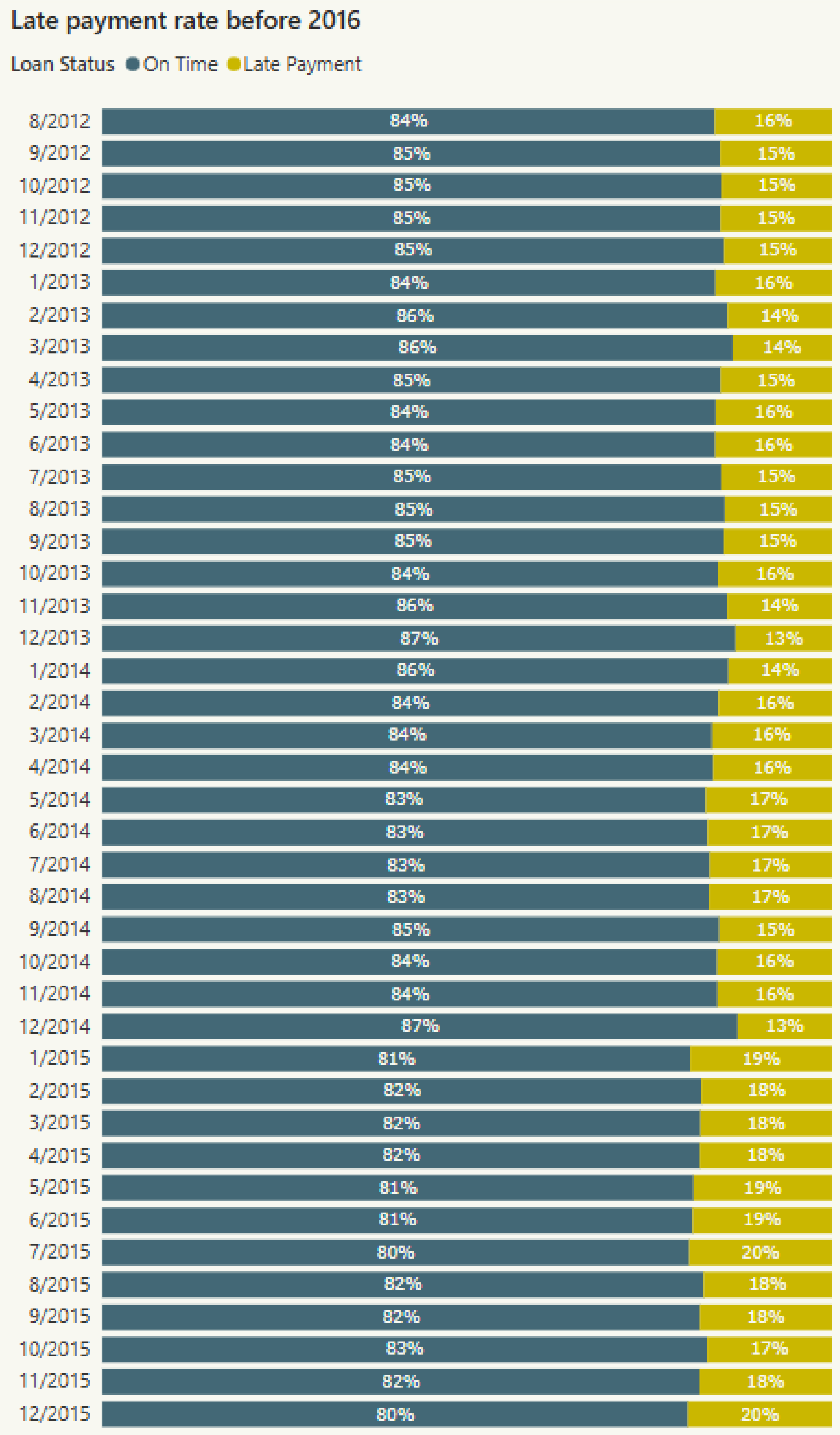
- ✓ Initial approach: Use the entire dataset to train and test model

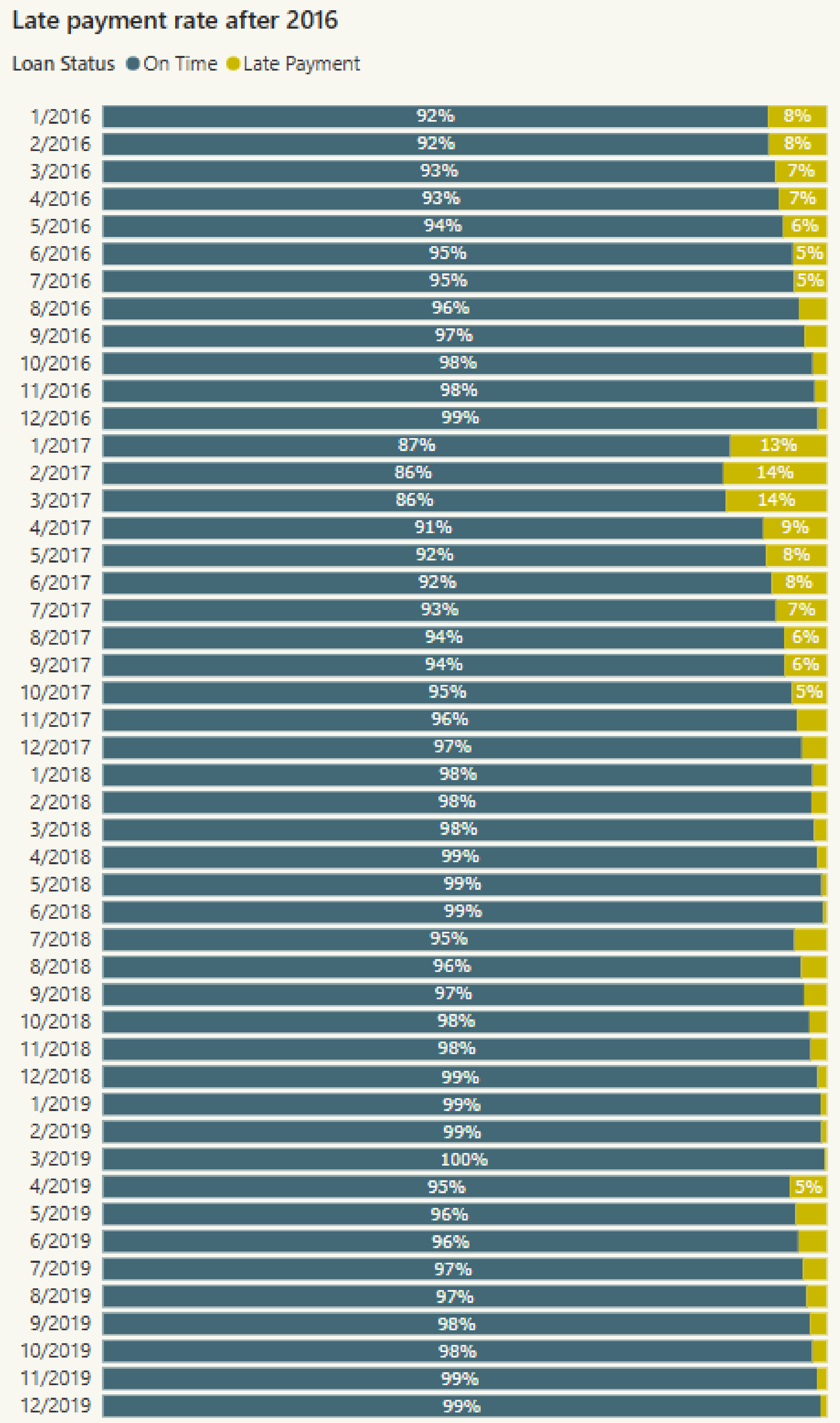
Complication:

- ✓ Data from 2016 onwards shows abnormal patterns: The late payment rate shows a gradual decline over an irregular cycle, followed by a sharp increase, and then another gradual decrease.
- ✓ Multiple data splitting methods were applied and models were trained using various approaches, but the model performance was extremely poor, with the recall rate not even reaching 50%.
- ✓ Unable to explain the cause of this anomaly.

Resolution:

- ✓ Limited data to 2012–2015, where:
  - Customer behavior patterns were stable
  - Model performance was strong and generalizable.





THANKS FOR

LISTENING