FPD Prediction Model Report

1. Overview

This report documents a high-level walkthrough the pipeline for building a machine learning model to predict First Payment Default (FPD) for loan applicants, using anonymized lead-level data. The goal is to help lenders assess the risk associated with each lead and potentially optimize decisioning strategies in real time.

2. Data Cleaning

2.1 Initial Inspection

The dataset was inspected for basic structure, missing values, and anomalies. There were a total of 58,801 rows of lead-level data with 399 columns, including a mix of ratio and count-based (for different entities such as phone, SSN, bank account, driving license, etc.) features aggregated across different time windows. There were some fields that had constant values throughout the entire dataset, hence they were dropped.

2.2 Handling Missing Data

- Binary missing indicators were retained for some fields (e.g., `lb_months_at_address_missing_True`).
- For categorical fields, an additional category "Missing" was created
- For numeric fields with few missing values, mean/median imputation was used (eg: months_at_address).
- Fields denoting "day_first_seen" or "day_last_seen" for an entity in the system had the highest % of missing values (as evident in Table 1). There were other features denoting lead date and # of days since first/last seen. Hence, these features were dropped.

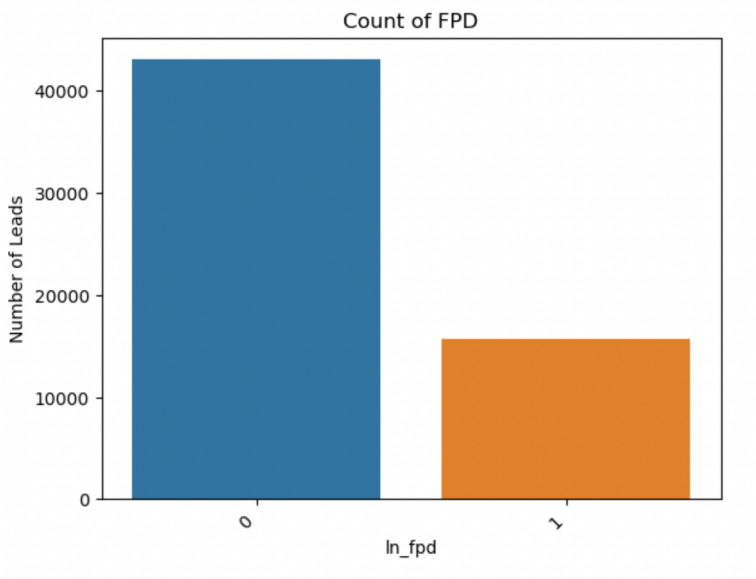
	Field Name	Field Type	# Records Have Values	% Populated	# Zeros	# Unique Values	Most Common
0	lead_datetime	categorical	58801	100.0%	314	628	2024-09-04 00:00:00
1	VarianceTable_variance_table_address_zip_valid	categorical	58801	100.0%	58801	1	TRUE
2	VarianceTable_variance_table_address_state	categorical	58801	100.0%	53515	2	Seen
3	VarianceTable_variance_table_address_zip	categorical	58801	100.0%	51665	2	Seen
4	VarianceTable_variance_table_address_city	categorical	58801	100.0%	47490	3	Seen
5	VarianceTable_variance_table_address_address	categorical	58801	100.0%	24517	4	New
6	VarianceTable_variance_table_bank_day_first_seen	categorical	48523	82.5%	85	2146	2023-05-31
7	VarianceTable_variance_table_bank_day_last_seen	categorical	48523	82.5%	168	1590	2024-08-30
8	VarianceTable_variance_table_bank_bank_account	categorical	58801	100.0%	40293	4	Match_Recent
9	VarianceTable_variance_table_bank_bank_aba	categorical	58801	100.0%	45183	3	Match_Recent
10	VarianceTable_variance_table_bank_account_type	categorical	58801	100.0%	53805	3	Match_Recent
11	VarianceTable_variance_table_bank_checksum	categorical	58801	100.0%	58801	1	TRUE
12	VarianceTable_variance_table_device_parent	categorical	48623	82.7%	19560	235	Mobile Safari Generic
13	VarianceTable_variance_table_device_platform	categorical	48623	82.7%	26559	13	iOS
14	VarianceTable_variance_table_device_browser	categorical	48623	82.7%	20977	29	Safari
15	VarianceTable_variance_table_device_device_type	categorical	48623	82.7%	26482	6	Mobile Device
16	VarianceTable_variance_table_device_device_pointing_method	categorical	48623	82.7%	42198	3	touchscreen
17	VarianceTable_variance_table_device_is_mobile_device	categorical	48623	82.7%	42198	2	TRUE
18	VarianceTable_variance_table_device_is_tablet	categorical	48623	82.7%	48390	2	FALSE
19	VarianceTable_variance_table_device_crawler	categorical	48623	82.7%	48618	2	FALSE
20	VarianceTable_variance_table_dob_dob	categorical	58801	100.0%	53599	2	Seen
21	VarianceTable_variance_table_driver_license_is_valid	categorical	50834	86.5%	44407	2	TRUE
22	VarianceTable_variance_table_driver_license_driver_license	categorical	58801	100.0%	39611	3	Seen
23	VarianceTable_variance_table_email_email	categorical	58801	100.0%	48583	4	Match_Recent
24	VarianceTable_variance_table_email_extension	categorical	58801	100.0%	33564	935	gmail.com
25	VarianceTable_variance_table_email_day_first_seen	categorical	52580	89.4%	165	2146	2018-11-19
26	VarianceTable_variance_table_email_day_last_seen	categorical	52580	89.4%	181	1714	2024-08-30
27	VarianceTable_variance_table_employer_income	categorical	58801	100.0%	29097	3	Range_Match_Recent
28	VarianceTable_variance_table_employer_employer_name	categorical	58798	100.0%	28939	4	Match_Recent
29	VarianceTable_variance_table_employer_pay_frequency	categorical	58801	100.0%	41682	3	Match_Recent
30	VarianceTable_variance_table_employer_pay_type	categorical	58801	100.0%	53225	3	Match_Recent
31	VarianceTable_variance_table_ip_ip	categorical	58544	99.6%	32546	3	New
32	VarianceTable_variance_table_name_last_name	categorical	58801	100.0%	52322	3	Seen
33	VarianceTable_variance_table_name_first_name	categorical	58801	100.0%	53041	3	Seen
34	VarianceTable_variance_table_phone_day_first_seen	categorical	52985	90.1%	156	2147	2018-11-19
35	VarianceTable_variance_table_phone_day_last_seen	categorical	52985	90.1%	181	1686	2023-05-22
36	VarianceTable_variance_table_phone_npa_npx	categorical	58801	100.0%	58779	2	TRUE
37	VarianceTable_variance_table_phone_home_phone	categorical	56017	95.3%	47968	3	Match_Recent
38	VarianceTable_variance_table_phone_cell_phone	categorical	58801	100.0%	50007	3	Match_Recent
39	VarianceTable_variance_table_ssn_ssn	categorical	58801	100.0%	53523	4	Seen
40	VarianceTable_variance_table_ssn_day_first_seen	categorical	53996	91.8%	200	2145	2018-11-19
41	VarianceTable_variance_table_ssn_day_last_seen	categorical	53996	91.8%	181	1770	2024-08-30
42	VarianceTable_variance_table_phone_work_phone	categorical	55436	94.3%	35412	3	Match_Recent
43	PricingTool_predictions_min	categorical	58801	100.0%	41696	5	"100+"
44	PricingTool_predictions_max	categorical	58801	100.0%	19548	5	"10-25"
45	PricingTool_predictions_ranked	categorical	58801	100.0%	11924	107	["0-10", "10-25", "25-50", "50-100", "100+"]
46	PricingTool_predictions_min_max	categorical	58801	100.0%	18420	20	["100+", "10-25"]

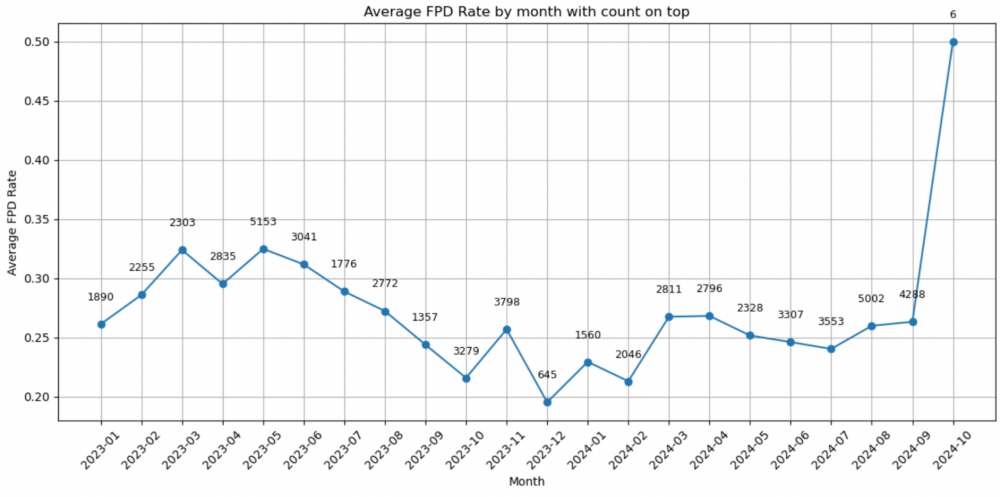
Table 1: Categorical Fields

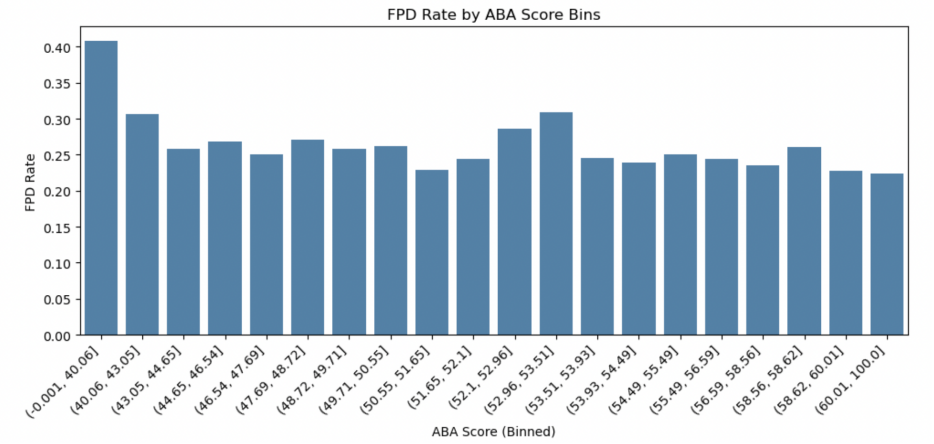
3. Exploratory Analysis

3.1 Target Distribution

- FPD was found to be imbalanced (~29% positive class), warranting the use of recall and AUC-ROC for evaluation



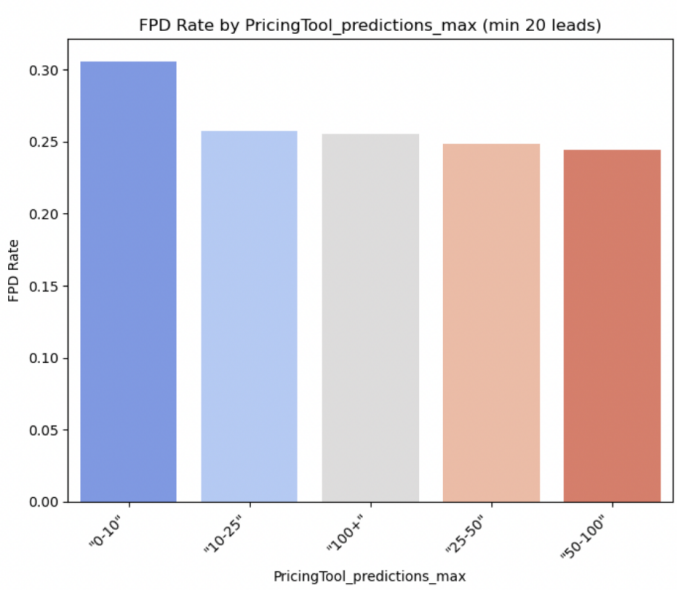


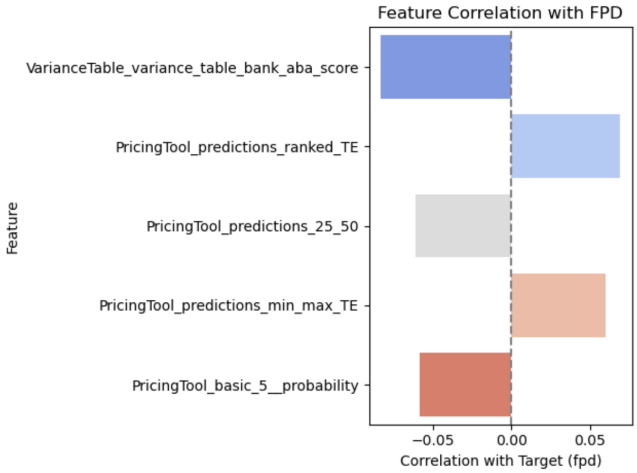


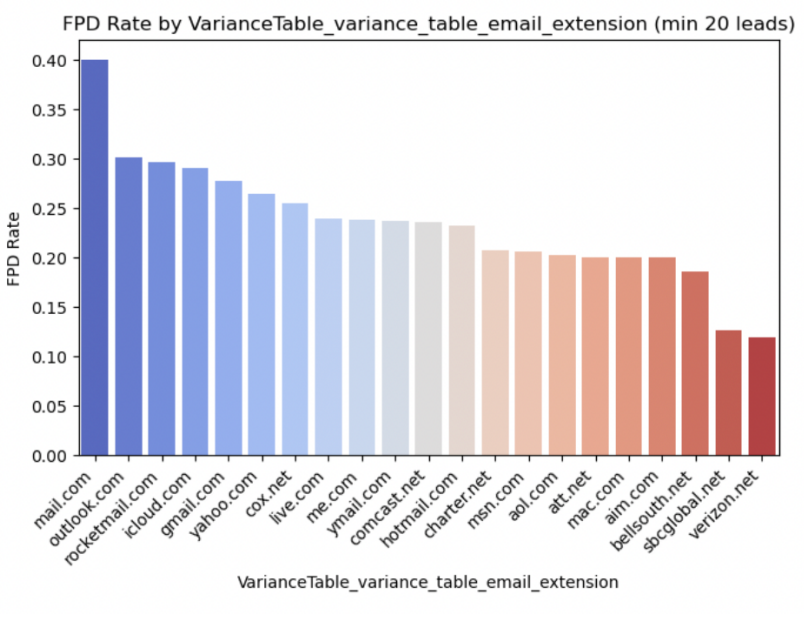
3.2 Feature-Level Insights

- Correlation among numerical variables was used to identify key variables. ABA Score, PricingTool projections, “days since first seen” measures and several SSN-based cross variances showed highest predictive power.

- Aggregated FPD rates across different categories of categorical variables provided insight into patterns that can be useful real-time detection. Fields such as email extension and parent device used showed higher FPD rates compared to their peer categories.







4. Feature Engineering

4.1 Numerical - Entity Stability and Behavioral Patterns

- Dataset already has aggregated features of time-windowed counts of appearances (1 day, 7 days, etc.). These did not need to be transformed and were directly leveraged to capture the volatility or stability of entities like SSN, phone, email.

- Ratio features denoting scale between appearances of an entity across different time windows offered insights into behavioral patterns. They were averaged across sets (6 hours and 24 hours for short term, 7 days and 30 days for mid-term and 90 days to all time for long term) to derive aggregated proxies for short-term,

mid-term and long-term behavior of applicant. Overall, these indicate relative activity change between time periods

- I also tried the same thing to compute deltas between time windows, but this led a very sparse matrix which did not add any incremental predictive power.

- Velocity features for all count-based fields to compute unusual activity. (velocity of x for 15 days = count_15_days/(average of count_15_days over past 15 days)

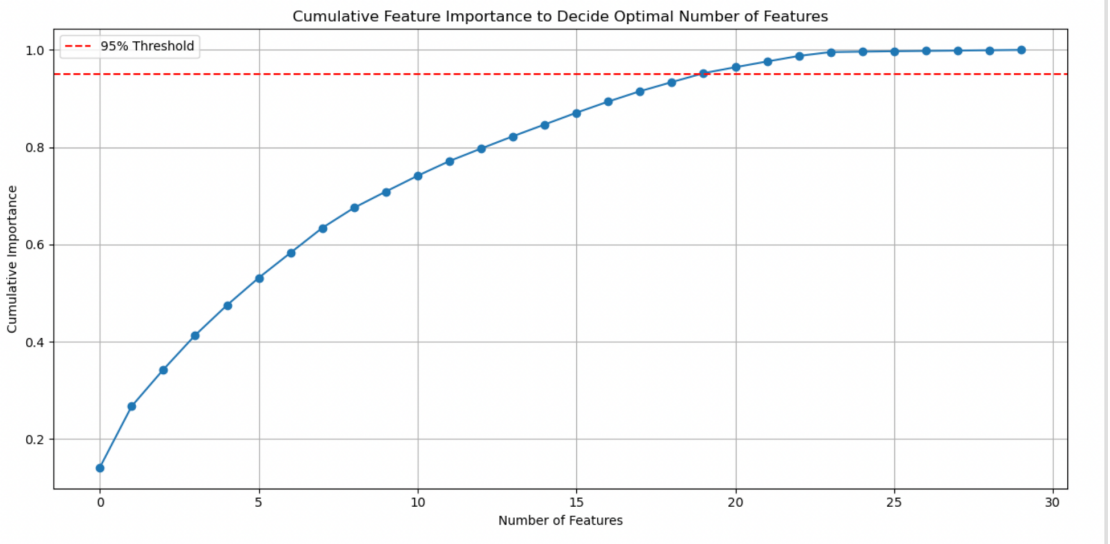
4.2 Categorical Features – Encoding

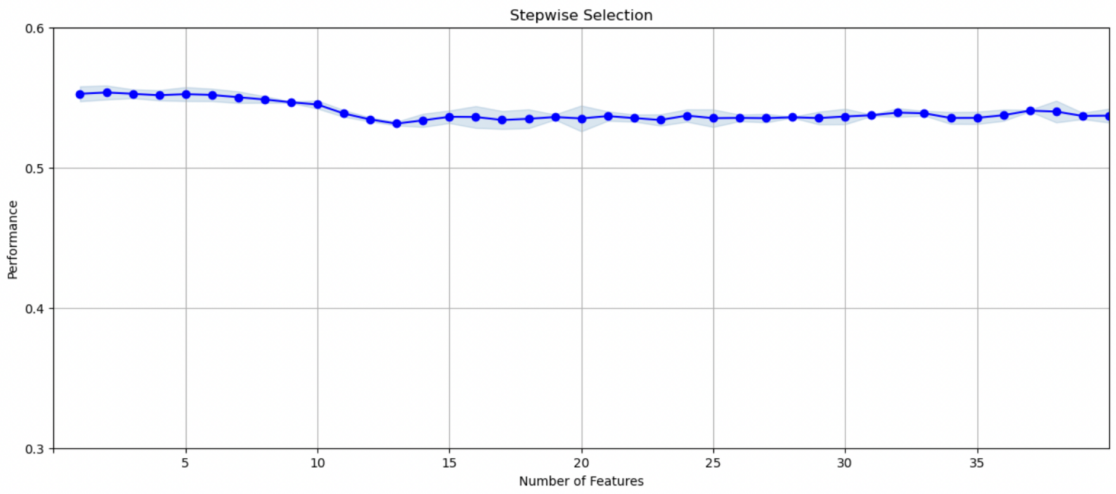
- Date & Time features were used to derive more features denoting the month, weekend or not?, etc.
- Categorical variables with high cardinality were target encoded with group mean to avoid extreme multidimensionality.
- Those with low cardinality were one-hot encoded with dummy variables.

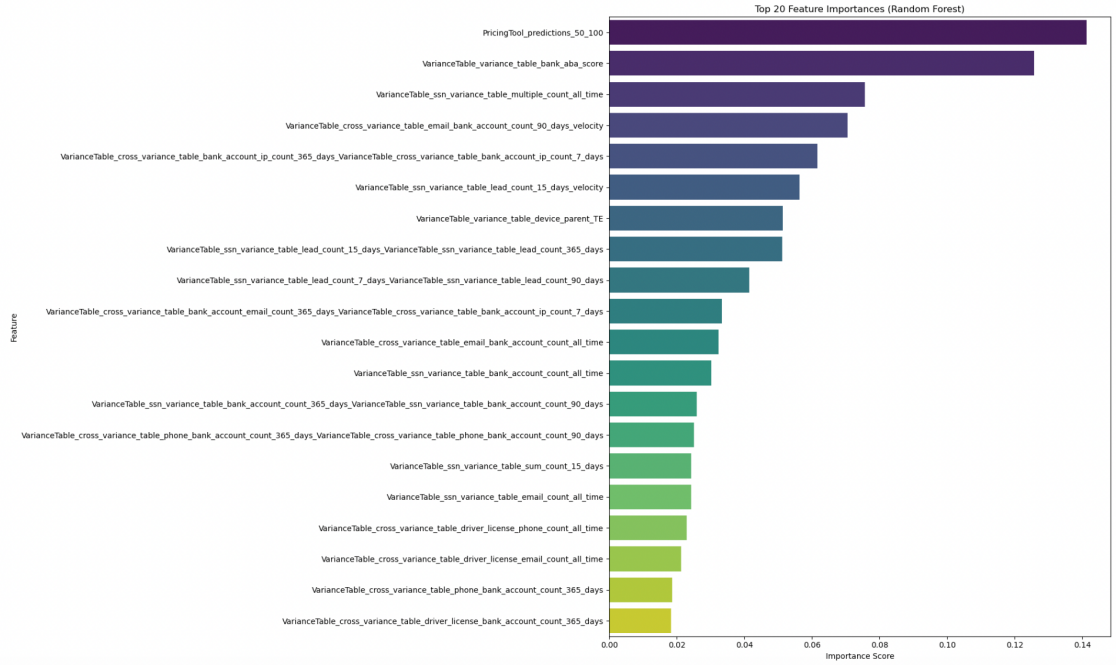
By the end of this process, there were a total of ~1300 columns.

5. Feature Selection

- Features with negligible variance were dropped.
- Used sequential feature selector to see number of features v/s performance (AUC)
- Used a simple random forest model to extract feature importances
- Constructed a cumulative feature importance plot to derive saturation point



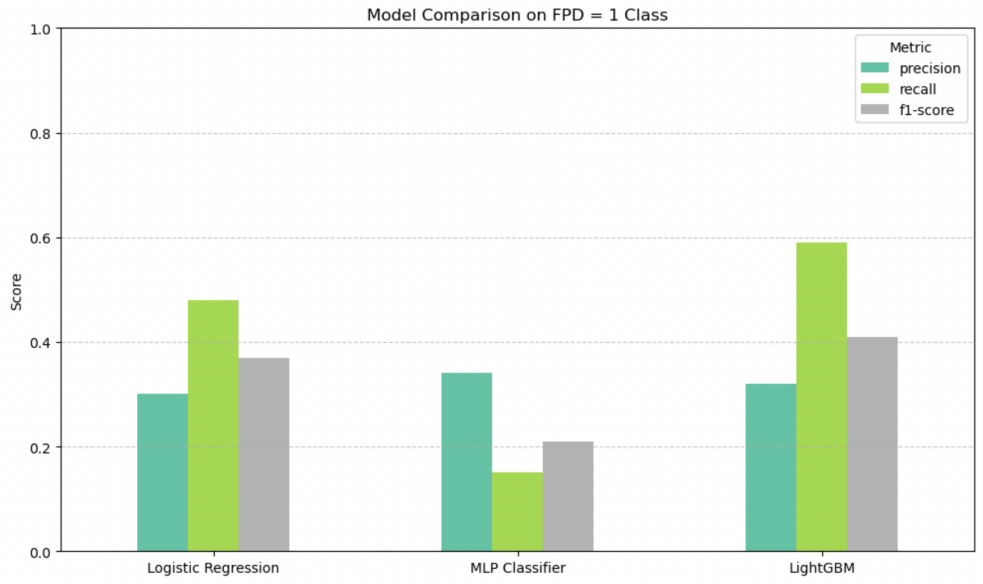




All of these steps led to ~30 features from the full list. Please refer to features_summary.csv file for final list of features with their descriptions.

6. Model Selection & Training

After splitting the data into training and test sets, I started off with a logistic regression model to establish a baseline and build upon it. I also experimented with KNN, Neural Network based MLP Classifier and Tree Based LightGBM model.



Ultimately, LightGBM Classifier was chosen due to better recall (and f-1 score). More generally, its speed and robustness with tabular data and imbalanced targets also contributed to choice of this model.

For training, I used GridSearchCV for fitting and tuning hyperparameters tuning with 5-fold cross validation (to avoid overfitting).

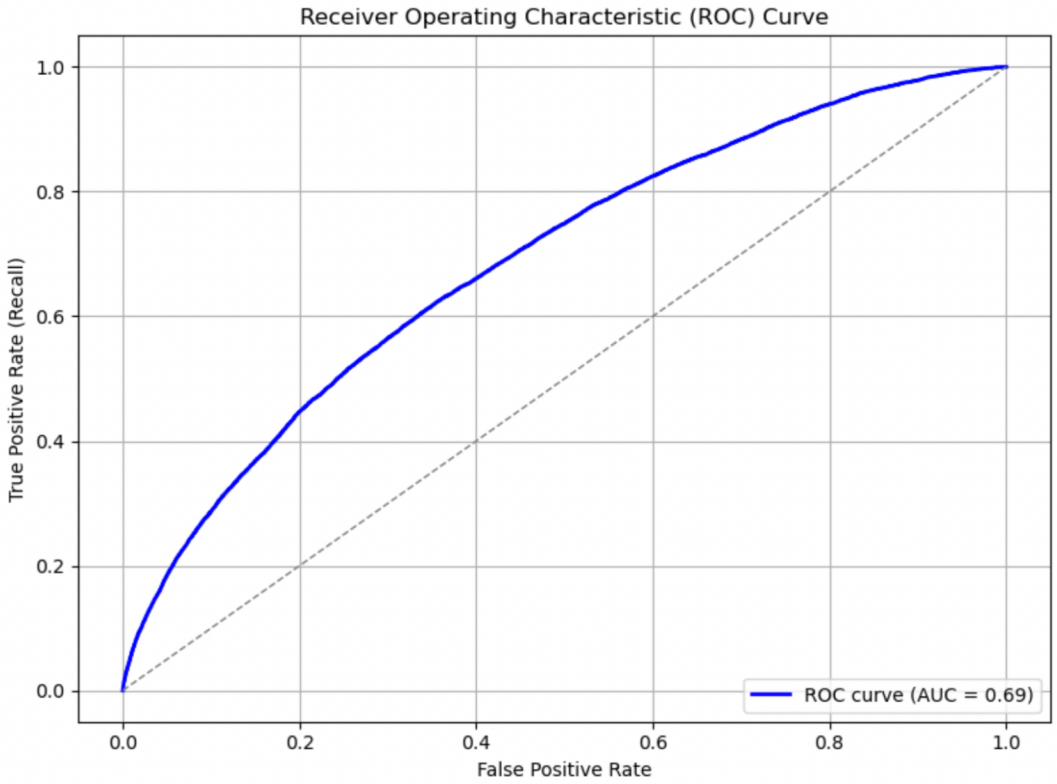
Final Hyperparameters:

- num_leaves=31
- max_depth=3
- col_sample_bytree = 0.8
- n_estimators=100
- scale_pos_weight=3
- subsample=0.6

Rest default values were used.

7. Model Evaluation

The model achieved an **AUC-ROC score of 0.7**, indicating a strong ability to distinguish between users who default and those who do not. For the positive class (i.e., users who failed to make the first payment), the model successfully identified **60% of actual defaulters** (recall) on **test set**. While the **precision was approximately 30%**, meaning that about one in three users flagged as defaulters were truly at risk, this trade-off was considered acceptable (depends upon lead acquisition cost vs loan payment). Accuracy was not the focus due to the natural class imbalance in the dataset, where defaulters are a minority.



The model can be used to inform strategic decisions such as lead prioritization or automatic filtering in loan disbursement workflows. By identifying applicants with a higher likelihood of first payment default, lenders can proactively manage risk and allocate resources more efficiently.

My recommendation to enhance the model's effectiveness would be that the model's probability thresholds should be tuned to support risk segmentation across low, medium, and high-risk borrower tiers. Retraining the model on a regular basis is recommended to capture evolving patterns, including potential seasonal effects. Finally, integrating the model into a real-time scoring pipeline will enable timely, interpretable decision-making at the point of application.

7. Future Scope

If I had more time to build this pipeline, I would try out the following options -

- **Advanced Feature Engineering:**
Go beyond raw variance fields and generate more granular features based on time-based decay functions.

- **External Data Integration:**
Integrate third-party data sources such as credit scores, outstanding current loans, loan terms, device risk scores, or geolocation risk profiles to enrich input features and add macroeconomic context if available.

- **Model Optimization & Stacking:**
Test a broader range of models (e.g., CatBoost, XGBoost, Naïve Bayes) and use ensemble methods or stacking to combine their strengths.

- **Custom Threshold Tuning:**
In our current pipeline, I have used a threshold of 0.5 (default value) to classify FPD. Experimenting with different thresholds according to different segments can really improve performance. Furthermore, using probability scores of the model to assign risk tiers can lead to optimal real-time predictions (low can be immediately approved, high can be rejected and medium can be flagged for manual review).