IISC Bengaluru CCE 2022

Computing for Artificial Intelligence and Machine Learning

Project Report

On

Credit Decision Model in Lending Industry

Predict the probability that customer does not pay back balance their credit card/ pay later card/ super card based on demographic & historical credit history data

Submitted by

Mr. Chetan Mehatre

Mr. Karan Rajput

Industry Overview -

The global FinTech lending market size was valued at $449.89 billion in 2020 and is projected to reach $4,957.16 billion by 2030, growing at a CAGR of 27.4% from 2021 to 2030. FinTech lender engages in using online technology to provide finance to businesses and individuals. Loans provided by these lenders are generally unsecured and take a personal guarantee from the borrowers. Moreover, it offers short-term loans with a quick turnaround from application to approval of loan amount. Furthermore, the FinTech lending process begins with online application with user registration, online documentation submission, customer authentication & verification, loan approval, loan disbursement, and loan recovery.

It allows borrowers to easily apply for loans and offers them transparency, which, in turn, leads to significant time savings. This automated nature of FinTech lending offers a more precise approach by delivering funding solutions, thus reducing stress and increasing chances of successful loan approval. However, rise in need for digital financing among individuals has led to increase in number of sophisticated cyberthreats where phishing scams, fraudulent behaviors, and malwares are improvising with increasing number of attempts. This is thereby becoming one of the major factors hampering the FinTech lending market growth.

On the contrary, implementation of cloud-based lending solutions enables FinTech firms to integrate supplementary infrastructure technologies with each other to create robust and highly secure platforms. This rising trend of cloud-based services is projected to provide lucrative opportunities for the FinTech lending market in the upcoming years.

Top Indian Companies in market -

- Jupiter Money
- Slice
- Fi Money
- Unicard
- Razorpay
- Navi

Facilities provided -

- Credit Card
- Buy Now Pay Later
- Super Card
- Personal Loans

Problem Statement:

- How to check credit worthiness of a customer?
- How much credit limit to be given?
- How to decide the interest rate for a particular customer?

Data Source -

American Express is a globally integrated payments company. The largest payment card issuer in the world, they provide customers with access to products, insights, and experiences that enrich lives and build business success.

Credit default prediction is central to managing risk in a consumer lending business. Credit default prediction allows lenders to optimize lending decisions, which leads to a better customer experience and sound business economics. Current models exist to help manage risk. But it's possible to create better models that can outperform those currently in use.

To predict the probability that a customer does not pay back their credit card balance amount in the future based on their monthly customer profile. The target binary variable is calculated by observing 18 months performance window after the latest credit card statement, and if the customer does not pay due amount in 120 days after their latest statement date it is considered a default event.

The dataset contains aggregated profile features for each customer at each statement date. Features are anonymized and normalized, and fall into the following general categories:

- D_* = Delinquency variables
- S_* = Spend variables
- P_* = Payment variables
- B_* = Balance variables
- R_* = Risk variables

Shape of data

Rows - 924621

Columns - 190

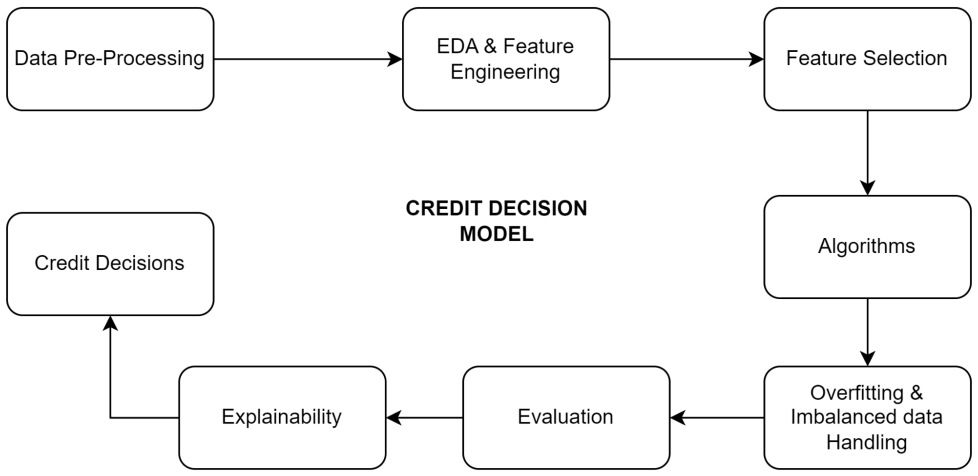
Performance window - 18 months

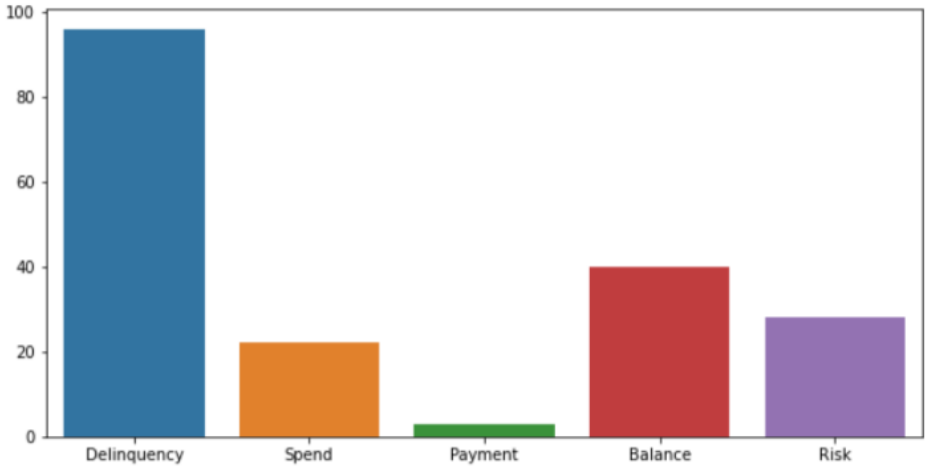
Target variable - 120+ DPD latest statement

Pre-Processing/Cleaning -

1. Stratified sample of 10000 data is chosen
2. Checking data info
3. Checking duplicates value
4. Checking Null Values
5. Dropped columns with 80% Null values
6. Filling remaining Null values with appropriate methods

Block Diagram:





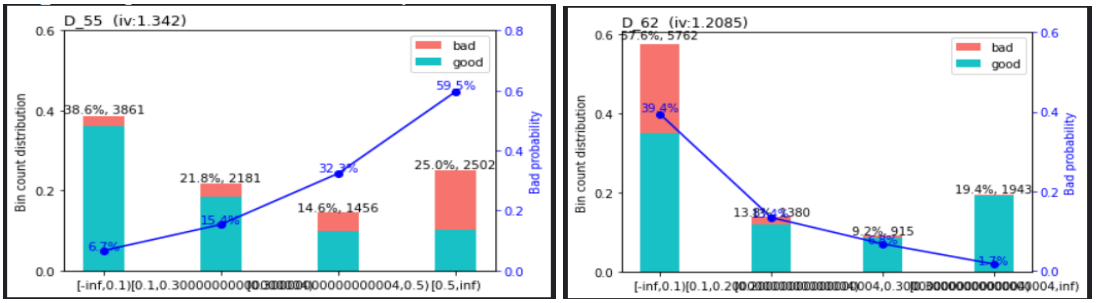
Feature selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables.

- Univariate Analysis
- Log Transformation
- Impute null values with appropriate methods
- Select top features based on feature importance values

Bi-Variate Plot: Trend Analysis



Distribution of features after selecting the top 29 features for further operations, this increases the speed, efficiency and accuracy of machine learning models

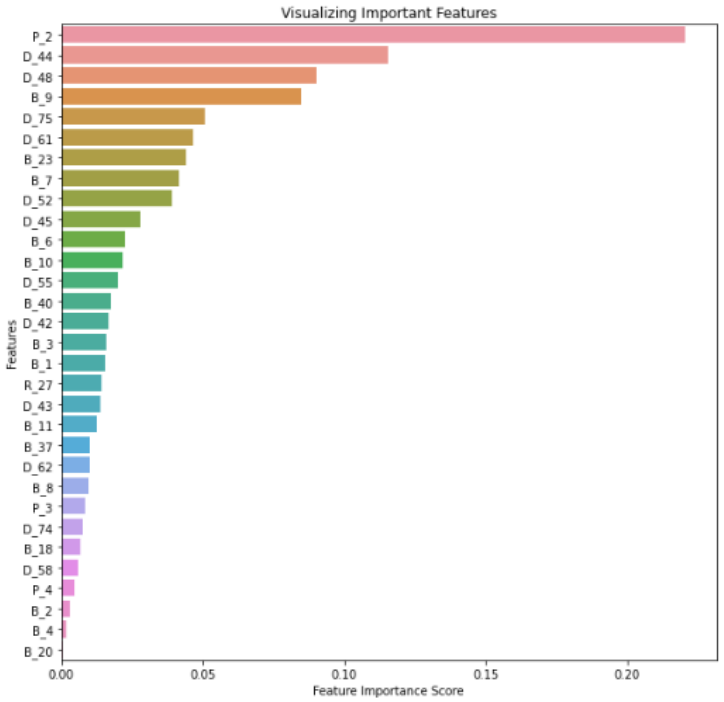


Fig. Feature importance of selected features

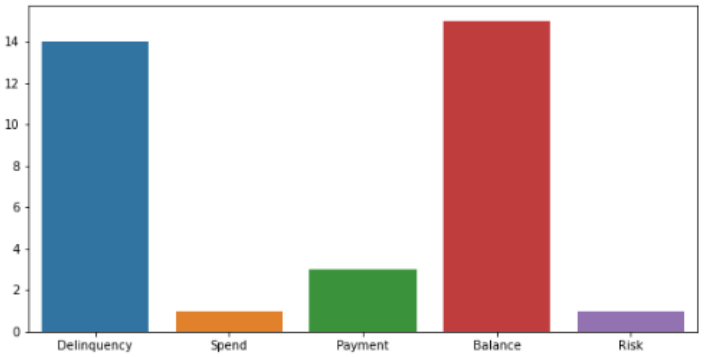


Fig. Distribution of selected features

Best fit Model:

Algorithms	ROC AUC	Kolmogorov-Smirnov Statistics	Gini
Logistic Regression	Train: 0.93 Test: 0.92	Train: 73 at decile 4 Test: 75 at decile 3	Train: 0.8 Test: 0.84
Random Forest	Train: 0.9469 Test: 0.9455	Train: 75.6 at decile 3 Test: 77.10 at decile 3	Train: 0.89 Test: 0.89
XGBoost	Train: 0.9767 Test: 0.9674	Train: 85.7 at decile 3 Test: 82.39 at decile 3	Train: 0.94 Test: 0.92

Kolmogorov-Smirnov Table:

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case).

Train:

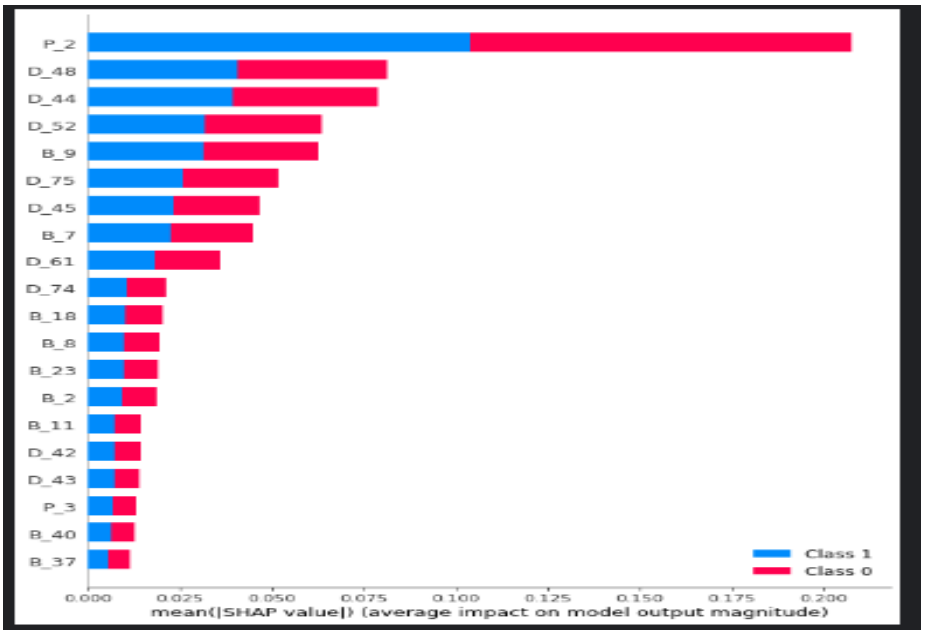
min_prob	max_prob	events	nonevents	event_rate	nonevent_rate	cum_eventrate	cum_noneventrate	KS	bad_rate
0.87	0.92	630	70	35.10%	1.34%	35.10%	1.34%	33.8	90.00%
0.78	0.87	546	154	30.42%	2.96%	65.52%	4.30%	61.2	78.00%
0.61	0.78	371	329	20.67%	6.32%	86.18%	10.62%	75.6	53.00%
0.39	0.61	180	520	10.03%	9.99%	96.21%	20.61%	75.6	25.71%
0.22	0.39	43	657	2.40%	12.62%	98.61%	33.24%	65.4	6.14%
0.12	0.22	11	689	0.61%	13.24%	99.22%	46.47%	52.7	1.57%
0.07	0.12	8	692	0.45%	13.29%	99.67%	59.77%	39.9	1.14%
0.05	0.07	5	695	0.28%	13.35%	99.94%	73.12%	26.8	0.71%
0.04	0.05	1	699	0.06%	13.43%	100.00%	86.55%	13.4	0.14%
0.04	0.04	0	700	0.00%	13.45%	100.00%	100.00%	0	0.00%

Test:

min_prob	max_prob	events	nonevents	event_rate	nonevent_rate	cum_eventrate	cum_noneventrate	KS	bad_rate
0.86	0.93	271	29	35.75%	1.29%	35.75%	1.29%	34.5	90.33%
0.77	0.86	221	79	29.16%	3.52%	64.91%	4.82%	60.1	73.67%
0.60	0.77	172	128	22.69%	5.71%	87.60%	10.53%	77.1	57.33%
0.38	0.60	61	239	8.05%	10.66%	95.65%	21.19%	74.5	20.33%
0.22	0.38	18	282	2.37%	12.58%	98.02%	33.76%	64.3	6.00%
0.12	0.22	5	295	0.66%	13.16%	98.68%	46.92%	51.8	1.67%
0.07	0.12	5	295	0.66%	13.16%	99.34%	60.08%	39.3	1.67%
0.05	0.07	5	295	0.66%	13.16%	100.00%	73.24%	26.8	1.67%
0.04	0.05	0	300	0.00%	13.38%	100.00%	86.62%	13.4	0.00%
0.04	0.04	0	300	0.00%	13.38%	100.00%	100.00%	0	0.00%

Explainability: Shapley Values

Shapley values are a widely used approach from cooperative game theory that come with desirable properties. The feature values of a data instance act as players in a coalition. The Shapley value is the average marginal contribution of a feature value across all possible coalitions.



Credit Risk Decisions - Business Outcome:

- Decide threshold to Approve/Reject customers
- Dynamic Limit Assignent
- Increase / Decrease limit / block card
- Risk based pricing
- Customer segmentation & Portfolio analysis
- Spending behaviour analysis