Loan Default Prediction

Model Report

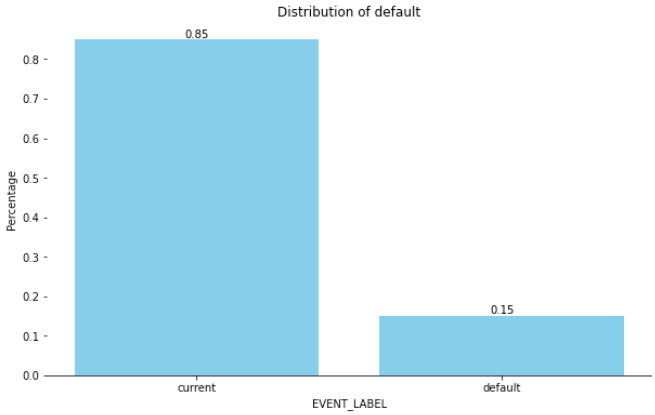
Name: Hewei Shen
Email: shenh23@wfu.edu

Part 1. Data Exploration and Preprocessing

1.1. Exploratory Data Analysis & Feature Screening

The raw dataset consists of 29777 rows and 52 columns. We selected 15 numeric and 10 categorical variables that can potentially be useful for default prediction to put in our model. See the Appendix for the data dictionary and missing values.

Our **target variable is the loan_status,** a binary variable with the value 'current' or 'default'. The distribution shows an imbalanced dataset with only 15.04% default transactions.

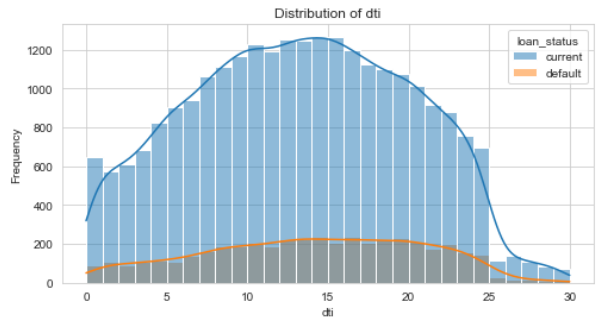


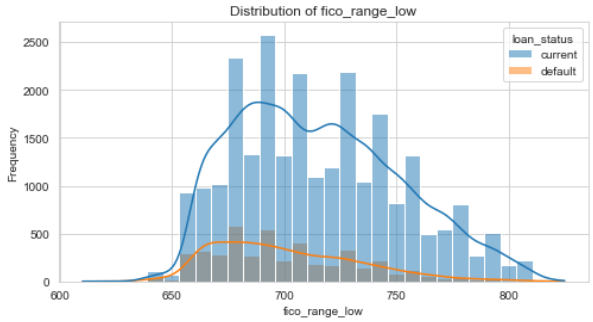
EDA on Numeric Variables

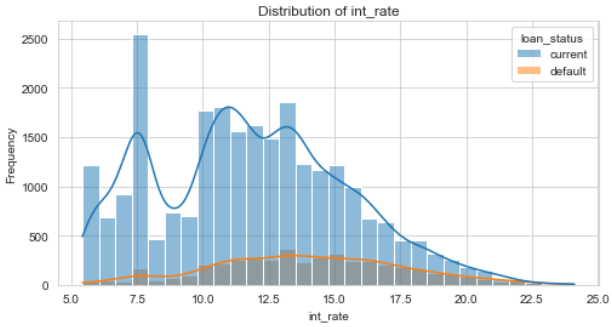
Exploratory analysis of numerical variables shows that a default account tends to have a larger debt-to-income ratio, a lower FICO score, a higher interest rate on the loan, a higher revolving line utilization rate, and a nearer date for the last credit record.

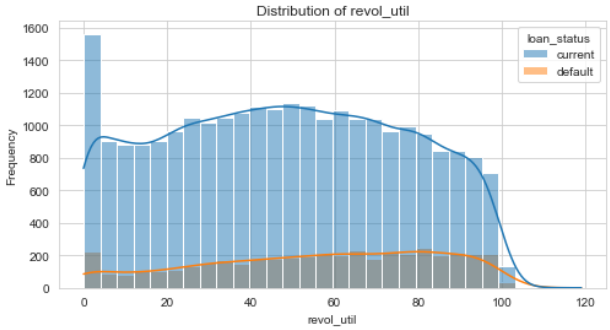
Outliers exist mostly in the variable of self-reported annual income, times of inquiry in the last 6 months, number of derogatory public records, number of public bankruptcy records, and date for the credit pullout last time.

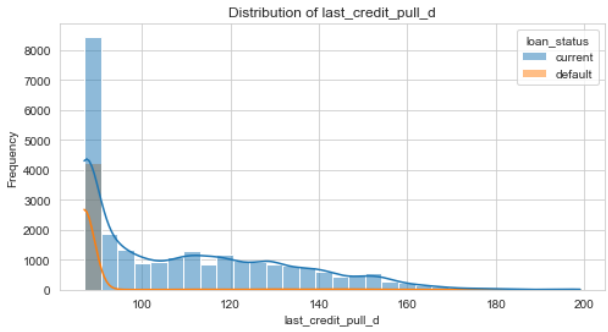
See the Appendix for a complete list of distribution graphs.



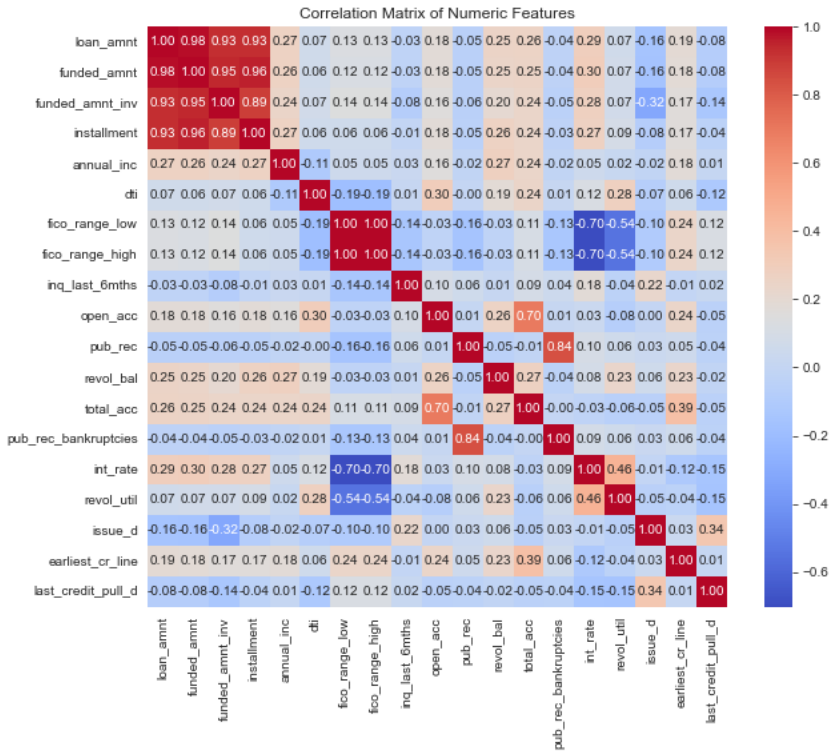








The correlation matrix of numeric features shows that some features share a high correlation.



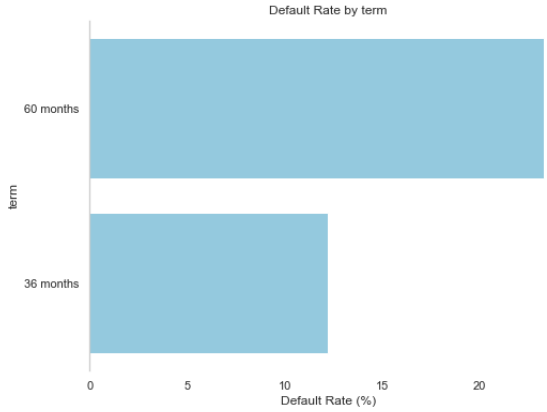
The loan amount applied by the borrower, the funded amount, the amount committed by investors, and the installment amount have a correlation coefficient of above 0.9 with each other. To avoid multicollinearity, we only chose the **installment amount** to put into the model. Similarly, we chose **fico_score_low** to put into the model and abandoned fico_score_high.

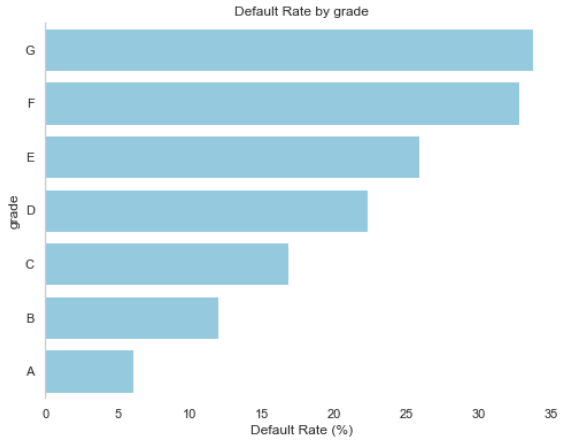
From the correlation matrix, we can tell that the revolving line utilization rate is one determinant factor of the FICO score, and the interest rate of the loan might largely be determined by the FICO score.

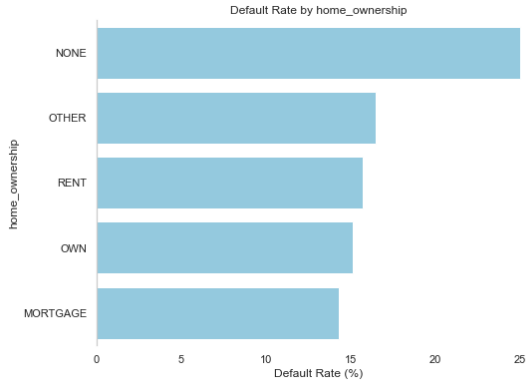
EDA on Categorical Variables

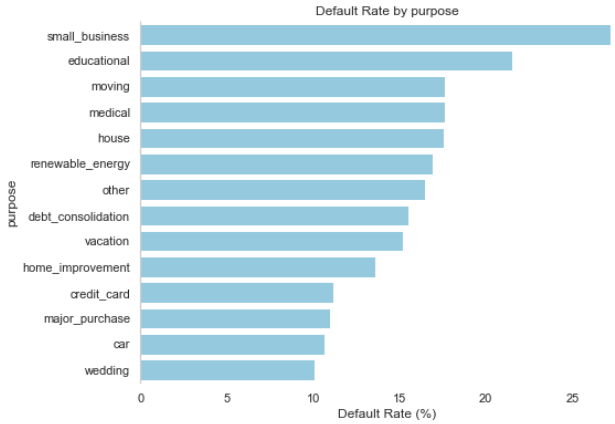
Exploratory analysis of categorical variables shows that a default account tends to have a longer loan term (in this case 60 months rather than 36 months), lower LC-assigned loan grade, home ownership type as ‘NONE’, and loan purpose as ‘small business’ and ‘educational’.

See the Appendix for a complete list of distribution graphs.









1.2. Data Preprocessing

Numerical variable

- Fill missing values in each column with the median value of that column
- Standardize features by removing the mean and scaling to unit variance, making the data normally distributed

Categorical variable

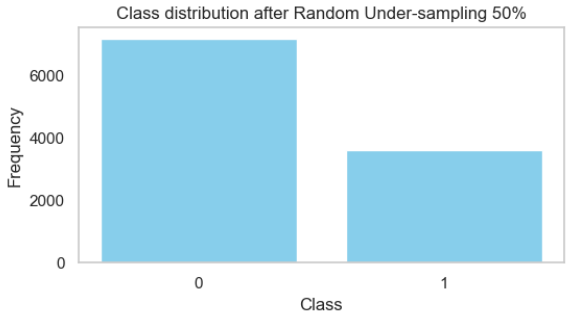
- Put missing values in the category ‘missing’ in each column
- Convert categorical variables into a matrix of binary variables, ignoring categories that were not seen during training

Part 2. Model Development

2.1. Deal with imbalanced data

Our dataset is imbalanced with the majority of loan_status value being 'current'. Imbalanced datasets can lead to poor model performance, because models trained on such data may become biased towards the majority class, often at the expense of the minority class's predictive accuracy.

We used the **under-sampling** technique to address the issue of imbalanced data. It selected a random sample of the majority class, in our case, the current transactions, to balance the weight of the majority and minority.



2.2. Model Training

We developed 5 classification models using **Logistic Regression, Random Forest, Gradient Boost, Neural Network, and Stacking Classifier** and trained on the training dataset. The optimal parameters for each model are listed below.

Logistic Regression	**Random Forest**	**Gradient Boost**	**Neural Nerwork**	**Stacking Classifier**
C: 0.1 max_iter: 300 penalty: l2	min_samples_split: 10 n_estimators: 50	learning_rate: 0.1 n_estimators: 200	hidden_layer_sized: (20,10,) activation: 'relu' solver: 'sgd' max_iter: 300	Base_estimators: gbm, rf, lr Final_estimator: lr

2.3. Parameter Tuning

The hyperparameter tuning of the models is completed by **Grid Search and Random Search**, which will try all the combinations of the parameter matrix and find the best performer. We used Grid Search on Logistic Regression and Gradient Boost; We used Random Search on Random Forest and Neural Network. In the Stacking Classifier, we used the best parameters in each model without tuning.

The model is trained using **3-fold cross-validation** to minimize the influence of sample selection on the final result.

Part 3. Global Model Explanation

3.1. Performance Metrics

The table lists the AUC, precision, recall, and F1 scores for both the training and test datasets.

Model	AUC (Train)	Precision (Train)	Recall (Train)	F1 (Train)	AUC (Test)	Precision (Test)	Recall (Test)	F1 (Test)
Logistic Regression	0.869	0.734	0.672	0.702	0.829	0.429	0.608	0.503
Random Forest	0.999	0.998	0.943	0.970	0.866	0.509	0.464	0.486
Gradient Boost	0.929	0.727	0.834	0.777	**0.894**	0.440	0.780	0.563
Neural Nerwork	0.973	0.863	0.914	0.888	0.881	0.427	0.756	0.546
Stacking Classifier	0.952	0.785	0.839	0.811	0.891	0.463	0.731	0.567

The AUC (area under the ROC curve) helps assess the model's overall ability to distinguish between defaulting and non-defaulting loans; Precision indicates the accuracy of identifying defaulted loans; Recall represents the proportion of actual defaulted loans identified correctly; F1 score balances the precision and recall, offering a comprehensive measure of the model's performance in predicting loan defaults.

We used AUC to determine the best-performing model because it's comparable across different models, while the other metrics are not comparable under the default threshold.

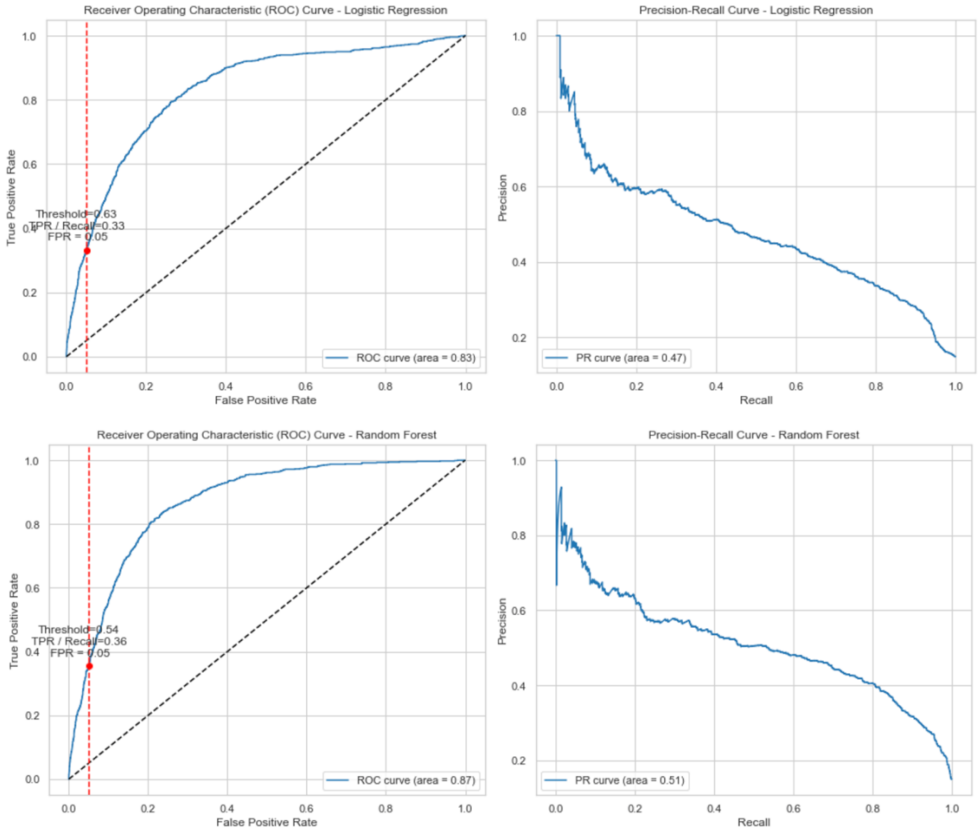
Based on the AUC score on the test set**, we chose Gradient Boost as our best model**, with an **AUC of 0.894.**

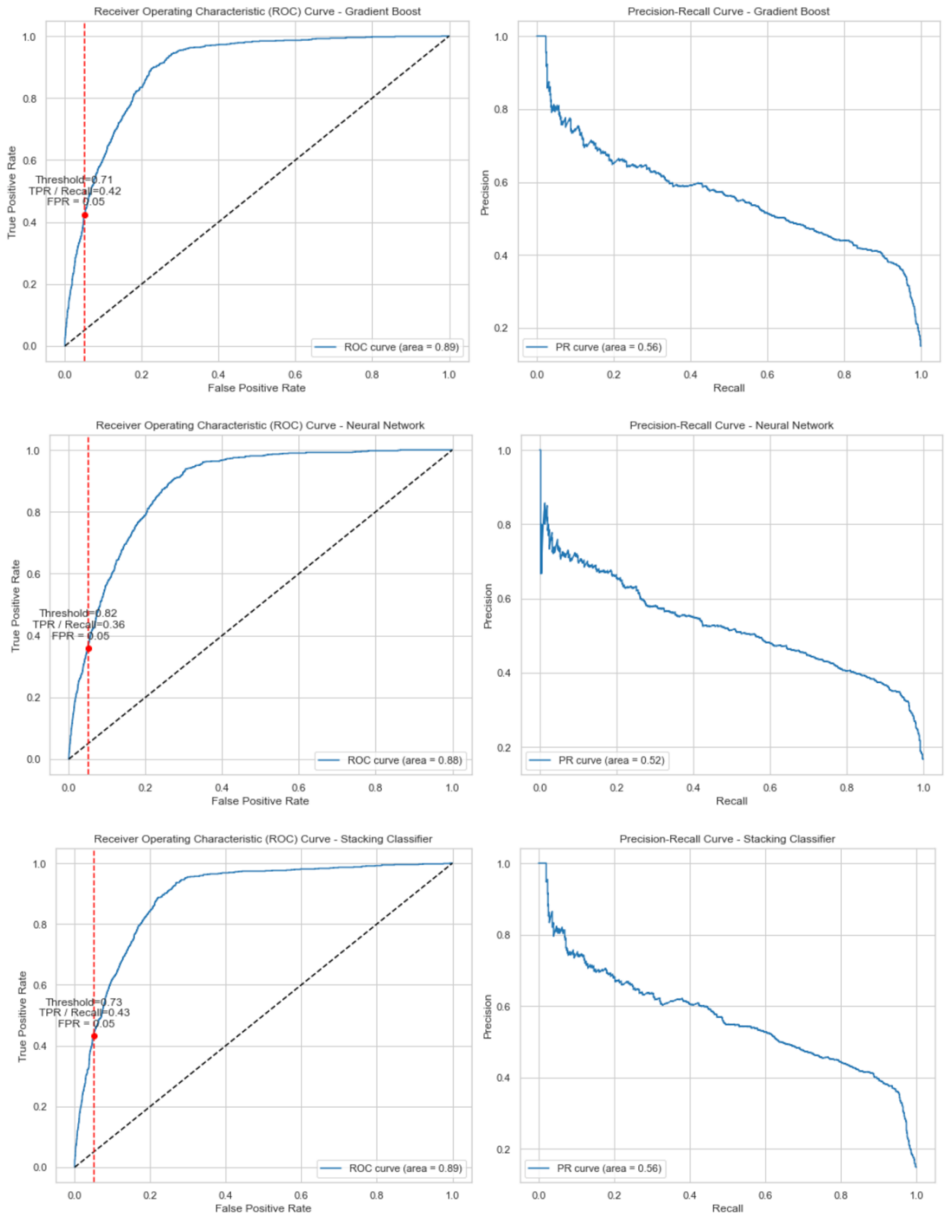
3.2. ROC and Precision-Recall Curve

The graphs show the ROC Curve and the Precision-Recall curve for each model. **Set the False Positive Rate (FPR) as 5%,** we can compare the True Positive Rate (TPR) / Recall across models and get the threshold for **achieving and maintaining the fixed FPR** (see the red dashed line).

The Precision-Recall Curve (PR Curve) shows a tradeoff between precision and recall. We can also choose to maintain a constant recall/precision level and set the threshold according to that.

See the Appendix for a complete list of operating tables and the ROC, PR Curve when FPR is set to 2%.





3.3. Feature Importance Analysis

In the graph, we listed the 10 most influential features in predicting default transactions according to each model using **permutation importance.**

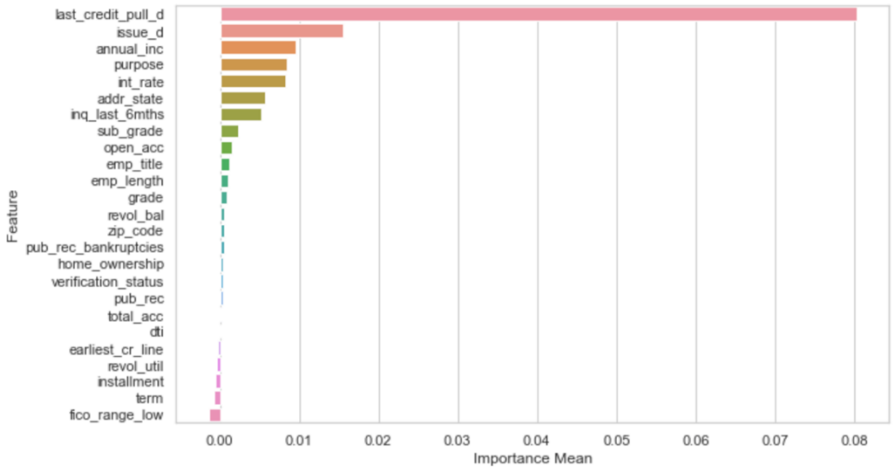
Permutation importance is a model-agnostic method used to assess the importance of features in machine learning models by evaluating how much the model's performance decreases when the values of a feature are randomly shuffled. This means that the values of that feature originally associated with each individual would be randomly reassigned to different individuals, but the

target variable remains the same for each data point. It provides insights into which features are most influential in making loan default predictions.

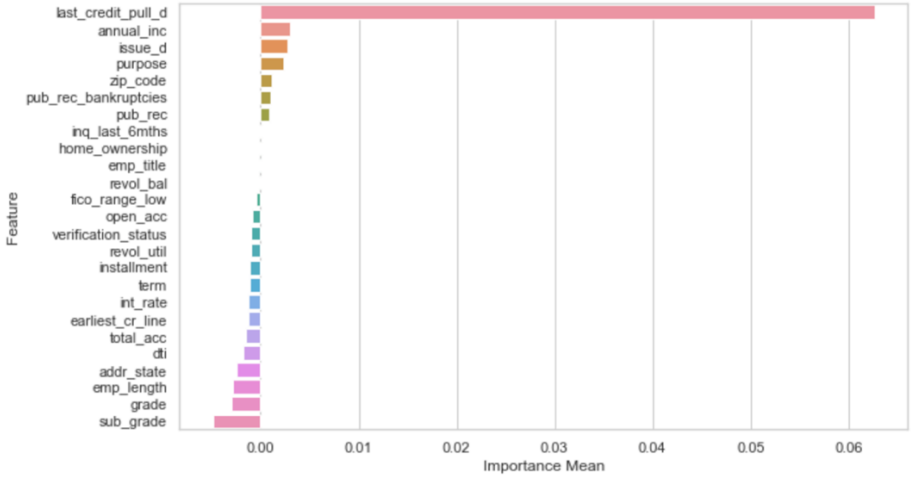
A positive importance score suggests that shuffling the values of that feature decreases the model's performance, indicating that the feature is important for predictions. Conversely, a negative importance score implies that shuffling improves the model's performance, suggesting that the feature may not be relevant or may even be harmful to the model's predictive ability.

The most important features detected by permutation importance show a commonality across different models. The last credit pull date, loan issue date, self-reported annual income, the interest rate of the loan, and the loan purpose are the **top 5 determinant factors** regarding default prediction. Among them, the **last credit pull date** has the most significant importance.

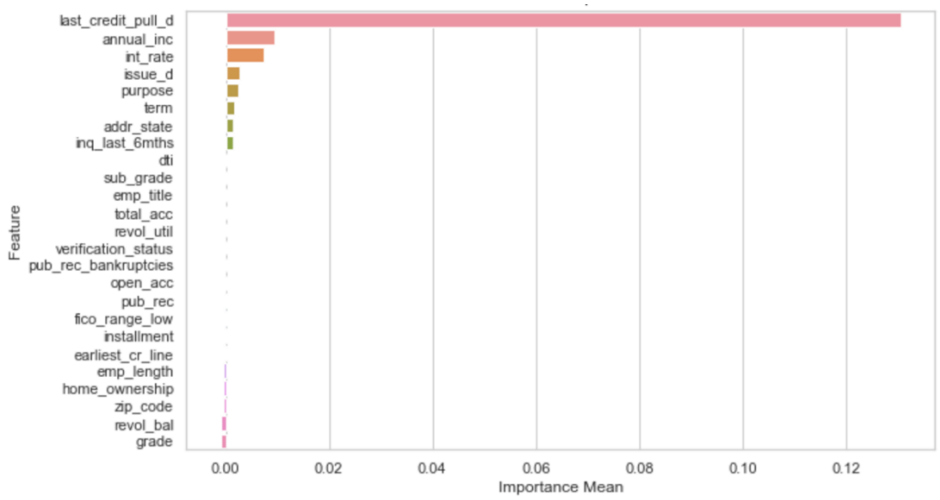
Feature Importance - Logistic Regression



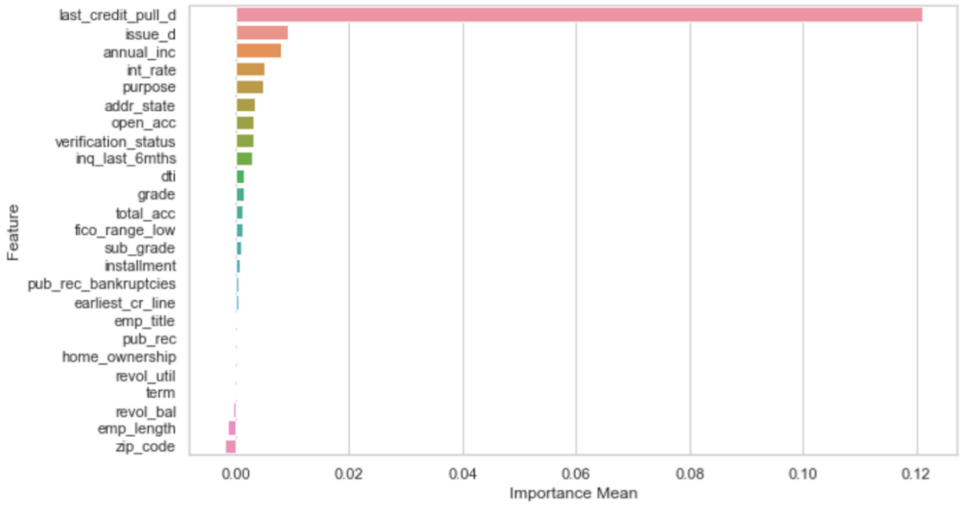
Feature Importance - Random Forest



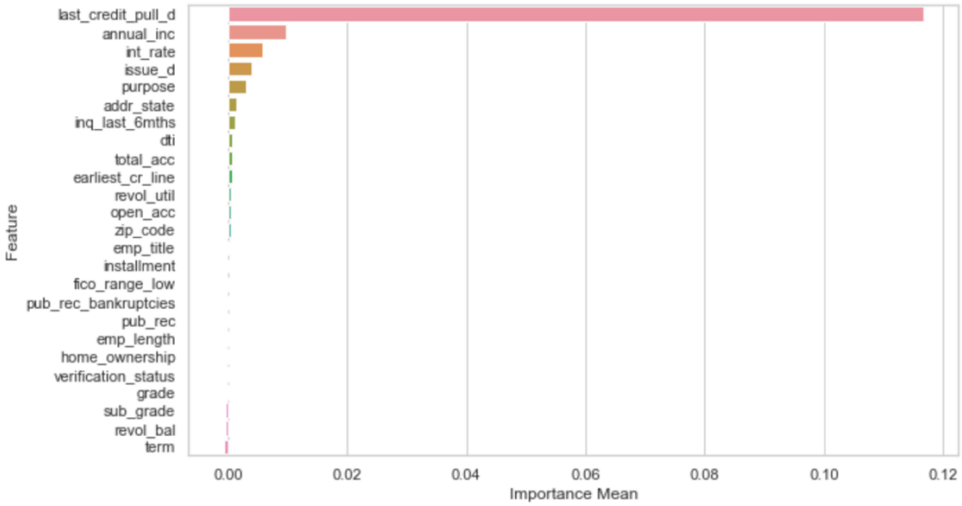
Feature Importance - Gradient Boost



Feature Importance - Neural Network



Feature Importance - Logistic Regression

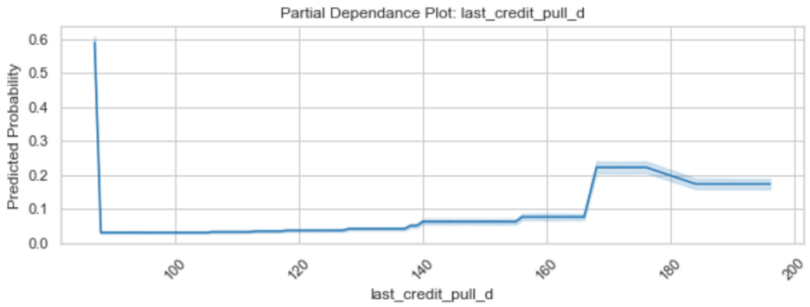


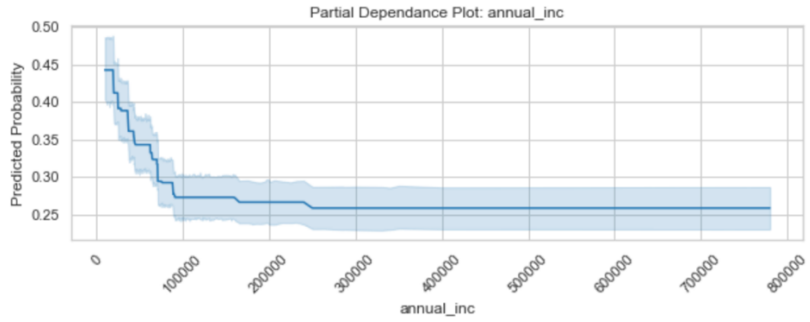
3.4. Partial Dependence Plot

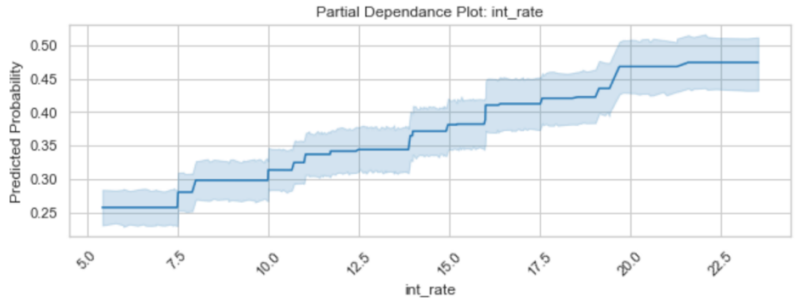
We draw the **Partial Dependence Plot (PDP) for the most important features** in our best model - **Gradient Boost** model.

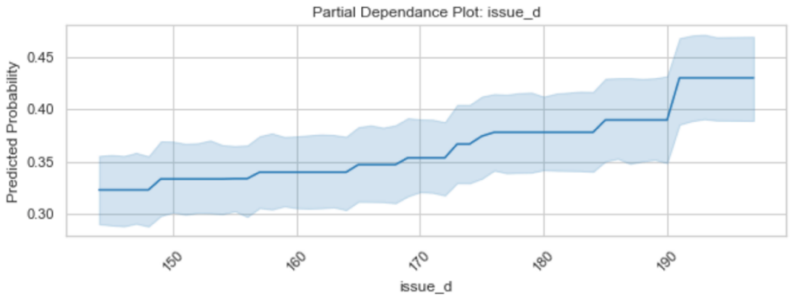
Partial dependence plots show the relationship between a feature and the predicted value while marginalizing the values of other features, providing insights into how the predicted default probability changes with variations in a specific feature, allowing us to understand the individual effect of a feature on the model's predictions.

For the last credit pull date and the issue date, a larger number means an earlier date. As shown in the graph. Borrowers with **a very near credit pull date of less than 90 and a further date larger than 170** have a higher probability of default.

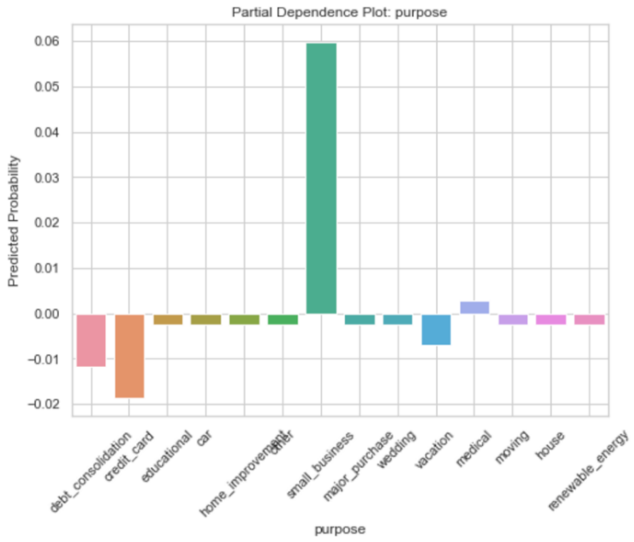








Borrowers with **an annual income of lower than 50,000, a higher interest rate, a further loan issue date, and a loan purpose of small-business funding** have a larger probability of default.



Part 4. Local Model Explanation

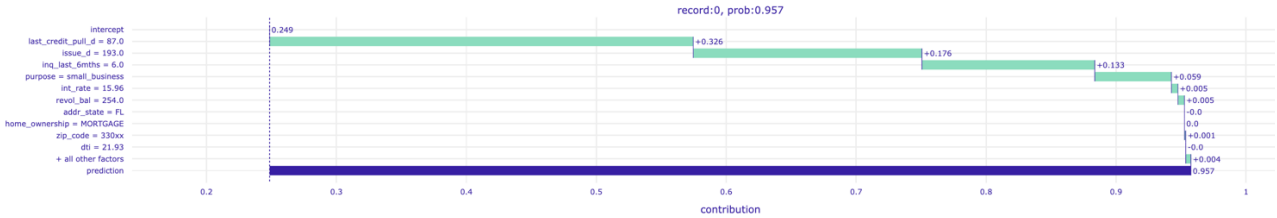
4.1 TOP 10 True Positives

This table listed the Top 10 True Positives predicted by the Gradient Boost model.

installment	annual_inc	dti	fico_range_low	inq_last_6mths	open_acc	pub_rec	revol_bal	total_acc	pub_rec_bankruptcies	...	emp_title	emp_length	home_ownership	purpose	zip_code	addr_state	verification_status	pred	pred_proba	loan_status
489.29	65000.0	21.93	655.0	6.0	9.0	0.0	254.0	30.0	NaN	...	Accushutters	1 year	MORTGAGE	small_business	330xx	FL	Not Verified	1	0.956980	1
795.11	616000.0	3.83	780.0	5.0	12.0	0.0	148829.0	43.0	NaN	...	SmartProperties.org Construction	10+ years	MORTGAGE	small_business	328xx	FL	Not Verified	1	0.950501	1
108.71	24000.0	2.00	660.0	0.0	3.0	0.0	469.0	7.0	0.0	...	Shetler Security Services	2 years	RENT	small_business	850xx	AZ	Source Verified	1	0.947289	1
296.46	85000.0	9.94	690.0	6.0	12.0	0.0	7491.0	15.0	0.0	...	NaN	< 1 year	RENT	small_business	916xx	CA	Verified	1	0.941127	1
57.41	35000.0	10.94	705.0	6.0	7.0	0.0	10008.0	12.0	NaN	...	bay area montesorri	< 1 year	RENT	medical	337xx	FL	Not Verified	1	0.935319	1
255.46	60000.0	3.96	765.0	5.0	4.0	0.0	7576.0	10.0	0.0	...	NaN	9 years	OWN	small_business	973xx	OR	Source Verified	1	0.931537	1
337.20	50000.0	18.77	735.0	2.0	2.0	0.0	517.0	16.0	0.0	...	Acapulco	2 years	RENT	small_business	917xx	CA	Source Verified	1	0.923515	1
235.33	14400.0	3.00	705.0	2.0	6.0	0.0	3449.0	6.0	0.0	...	NaN	2 years	RENT	other	937xx	CA	Not Verified	1	0.923440	1
196.18	7000.0	8.57	680.0	2.0	1.0	0.0	0.0	2.0	NaN	...	UWF Parking Services	< 1 year	RENT	house	325xx	FL	Not Verified	1	0.921421	1
268.95	13000.0	0.00	715.0	3.0	5.0	0.0	0.0	5.0	0.0	...	Mainstay Business Solutions	< 1 year	RENT	educational	933xx	CA	Not Verified	1	0.917564	1

The graph provided a breakdown of the predicted default probability into each factor for the Top 1 True Positive. Each feature has an incremental contribution to the final probability. In this graph,

the last credit pull date, the issue date, and counts of inquiries in the last 6 months contribute most of the predicting power.

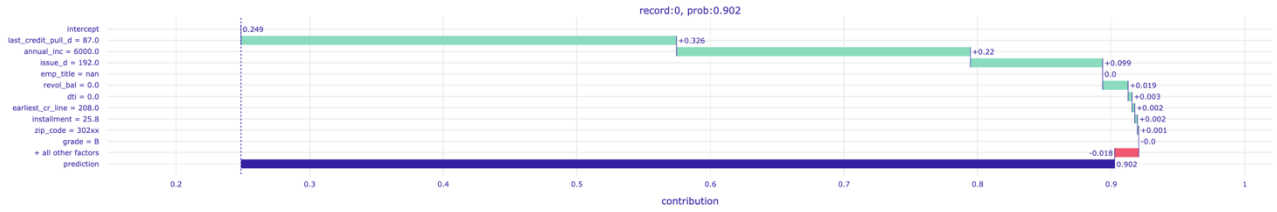


4.2 TOP 10 False Positives

This table listed the Top 10 False Positives predicted by the Gradient Boost model.

installment	annual_inc	dti	fico_range_low	inq_last_6mths	open_acc	pub_rec	revol_bal	total_acc	pub_rec_bankruptcies	...	emp_title	emp_length	home_ownership	purpose	zip_code	addr_state	verification_status	pred	pred_proba	loan_status
25.80	6000.0	0.00	680.0	1.0	5.0	0.0	0.0	5.0	NaN	...	NaN	< 1 year	RENT	debt_consolidation	302xx	GA	Not Verified	1	0.902170	0
207.22	20000.0	18.18	690.0	0.0	2.0	0.0	1315.0	11.0	0.0	...	home depot	3 years	RENT	debt_consolidation	935xx	CA	Verified	1	0.899252	0
46.91	17000.0	20.40	670.0	6.0	8.0	0.0	3368.0	8.0	0.0	...	University of Minnesota	2 years	RENT	major_purchase	557xx	MN	Not Verified	1	0.897506	0
69.32	49000.0	22.10	660.0	2.0	15.0	0.0	8158.0	34.0	0.0	...	Michael Enterprises	10+ years	MORTGAGE	small_business	497xx	MI	Not Verified	1	0.896989	0
133.81	36000.0	7.17	670.0	0.0	3.0	0.0	3471.0	9.0	0.0	...	bed bath and beyond	3 years	RENT	other	070xx	NJ	Verified	1	0.888755	0
34.59	18000.0	0.00	710.0	2.0	11.0	0.0	0.0	11.0	0.0	...	Aurora Multimedia	< 1 year	RENT	moving	088xx	NJ	Source Verified	1	0.884852	0
186.46	36000.0	16.03	650.0	4.0	7.0	0.0	3132.0	20.0	NaN	...	Wisconsin Business Development Finance Corpora...	< 1 year	RENT	debt_consolidation	532xx	WI	Not Verified	1	0.883408	0
437.92	42000.0	9.03	715.0	1.0	6.0	0.0	13399.0	12.0	0.0	...	Riverwind Casino	5 years	MORTGAGE	small_business	731xx	OK	Verified	1	0.883073	0
139.17	10800.0	0.00	745.0	2.0	3.0	0.0	0.0	10.0	0.0	...	Wells Fargo	< 1 year	RENT	moving	941xx	CA	Not Verified	1	0.882881	0
268.95	81600.0	4.97	785.0	3.0	4.0	0.0	11.0	26.0	0.0	...	NaN	5 years	RENT	small_business	605xx	IL	Verified	1	0.879306	0

The graph provided a breakdown of the predicted default probability into each factor for the Top 1 False Positive.



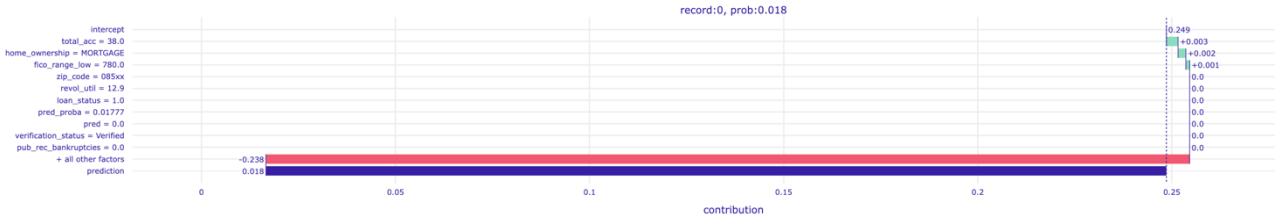
A common theme for the Top 10 False Positives is that their probability of default is determined by the most important factors. When the value of the most important features falls into the groups that have the most determinant power, while they are actually the rare cases that are non-default, there will be False Positives.

4.3 TOP 10 False Negatives

This table listed the Top 10 False Negatives predicted by the Gradient Boost model.

installment	annual_inc	dti	fico_range_low	inq_last_6mths	open_acc	pub_rec	revol_bal	total_acc	pub_rec_bankruptcies	...	emp_title	emp_length	home_ownership	purpose	zip_code	addr_state	verification_status	pred	pred_proba	loan_status
701.48	98000.0	20.85	780.0	0.0	20.0	0.0	12575.0	38.0	0.0	...	East Windsor Regional School District	10+ years	MORTGAGE	debt_consolidation	085xx	NJ	Verified	0	0.017775	1
311.02	59000.0	9.82	780.0	0.0	4.0	0.0	83.0	16.0	0.0	...	csc	10+ years	OWN	major_purchase	890xx	NV	Verified	0	0.021984	1
247.29	45500.0	18.78	680.0	0.0	16.0	0.0	10933.0	33.0	0.0	...	GEICO	5 years	MORTGAGE	credit_card	142xx	NY	Not Verified	0	0.023312	1
164.86	37000.0	19.20	670.0	0.0	10.0	0.0	5505.0	15.0	0.0	...	Macys	< 1 year	RENT	credit_card	236xx	VA	Verified	0	0.023676	1
233.06	150000.0	8.60	695.0	7.0	6.0	0.0	21293.0	9.0	0.0	...	The Perfect Body, Inc.	6 years	OWN	credit_card	327xx	FL	Not Verified	0	0.024803	1
152.17	120000.0	17.03	720.0	0.0	14.0	0.0	19237.0	23.0	0.0	...	ARINC	9 years	MORTGAGE	home_improvement	741xx	OK	Not Verified	0	0.024882	1
154.71	54000.0	10.71	765.0	1.0	15.0	0.0	3371.0	29.0	0.0	...	polk county school board	10+ years	MORTGAGE	home_improvement	338xx	FL	Not Verified	0	0.025856	1
252.93	110656.0	7.82	670.0	0.0	13.0	0.0	9869.0	34.0	0.0	...	OnLive Inc	3 years	MORTGAGE	major_purchase	940xx	CA	Source Verified	0	0.026989	1
94.82	42120.0	15.04	670.0	0.0	5.0	1.0	3266.0	16.0	1.0	...	St Johns med center	3 years	RENT	moving	930xx	CA	Not Verified	0	0.027312	1
104.75	84000.0	24.06	715.0	2.0	5.0	0.0	422.0	20.0	0.0	...	Brawley Insurance Services	3 years	RENT	major_purchase	937xx	CA	Source Verified	0	0.027315	1

The graph provided a breakdown of the predicted default probability into each factor for the Top 1 False Negative.



A common theme for the Top 10 False Negatives is that their probability of default is dragged down by all other factors that are not listed as the most important. When the value of the most important features doesn't fall into the groups that have the most determinant power, the other factors come out and bring uncertainty to the prediction.

(End of model report)

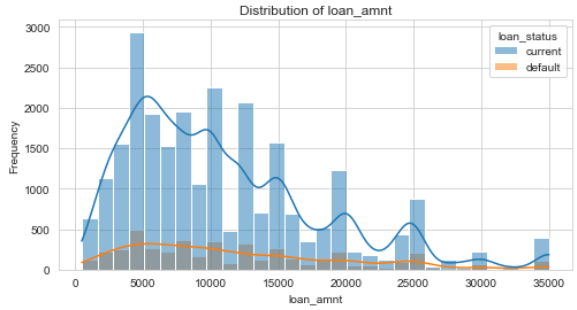
Appendix

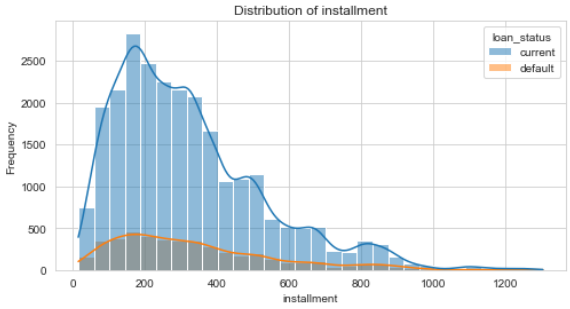
Missing Value in the Dataset

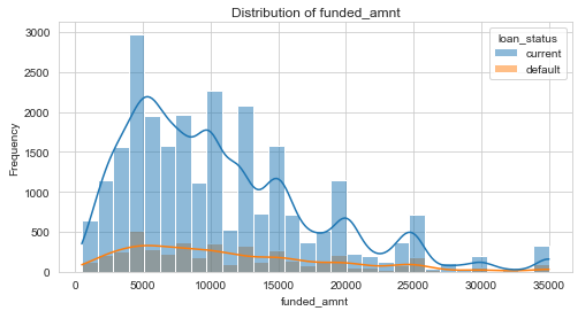
	Variable Name	Missing Percentage
0	id	0.01%
1	member_id	0.01%
2	loan_amnt	0.01%
3	funded_amnt	0.01%
4	funded_amnt_inv	0.01%
5	term	0.01%
6	int_rate	0.01%
7	installment	0.01%
8	grade	0.01%
9	sub_grade	0.01%
10	emp_title	6.12%
11	emp_length	2.56%
12	home_ownership	0.01%
13	annual_inc	0.01%
14	verification_status	0.01%
15	issue_d	0.01%
16	loan_status	0.00%
17	pymnt_plan	0.01%
18	url	0.01%
19	desc	31.68%
20	purpose	0.01%
21	title	0.05%
22	zip_code	0.01%
23	addr_state	0.01%
24	dti	0.01%
25	delinq_2yrs	0.08%
26	earliest_cr_line	0.08%
27	fico_range_low	0.01%
28	fico_range_high	0.01%
29	inq_last_6mths	0.08%
30	mths_since_last_delinq	63.50%
31	mths_since_last_record	91.37%
32	open_acc	0.08%
33	pub_rec	0.08%
34	revol_bal	0.01%
35	revol_util	0.23%
36	total_acc	0.08%
37	out_prncp	0.01%
38	out_prncp_inv	0.01%
39	total_rec_late_fee	0.01%
40	last_pymnt_d	0.23%
41	last_pymnt_amnt	0.01%

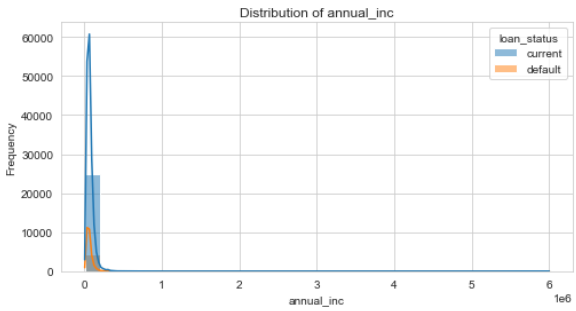
42	next_pymnt_d	92.10%
43	last_credit_pull_d	0.02%
44	collections_12_mths_ex_med	0.35%
45	policy_code	0.01%
46	application_type	0.01%
47	acc_now_delinq	0.08%
48	chargeoff_within_12_mths	0.35%
49	delinq_amnt	0.08%
50	pub_rec_bankruptcies	3.24%
51	tax_liens	0.27%

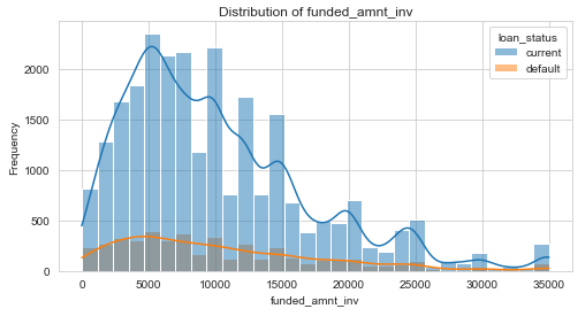
EDA on Numeric Features

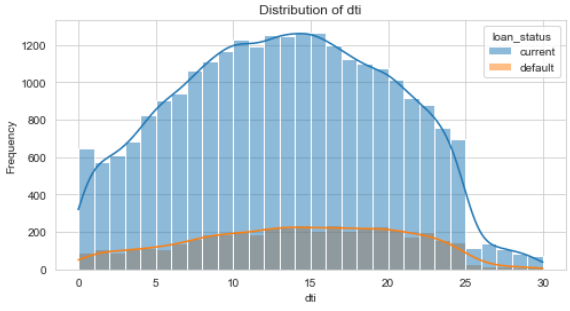


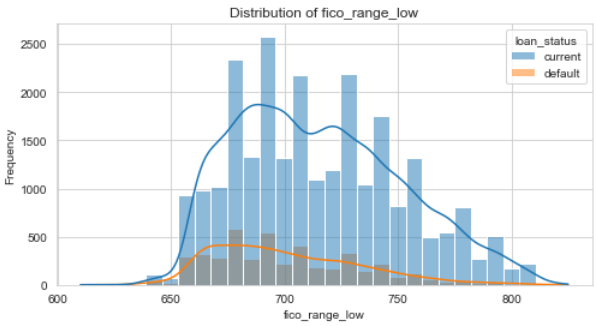


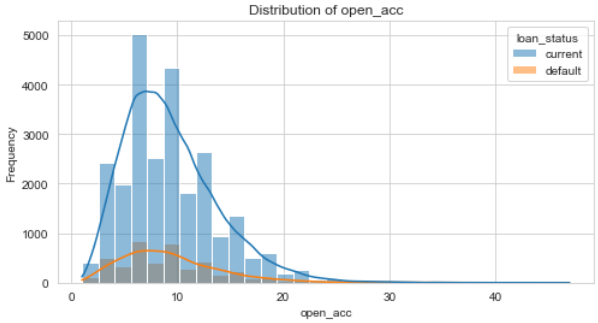


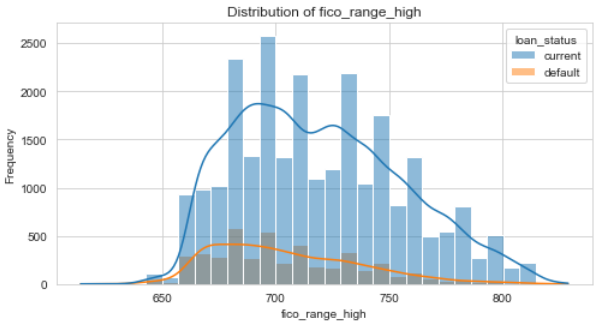


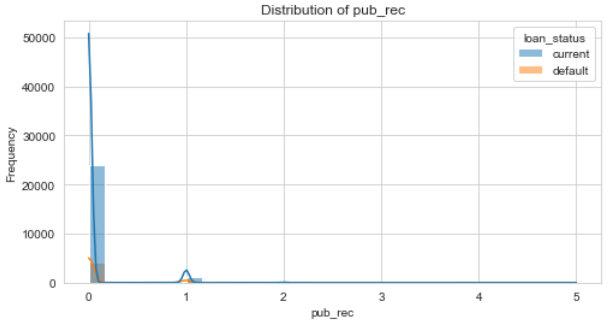


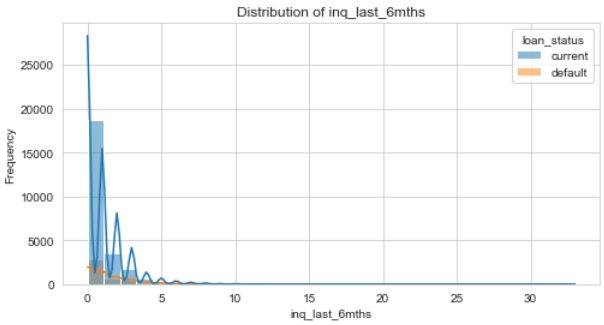


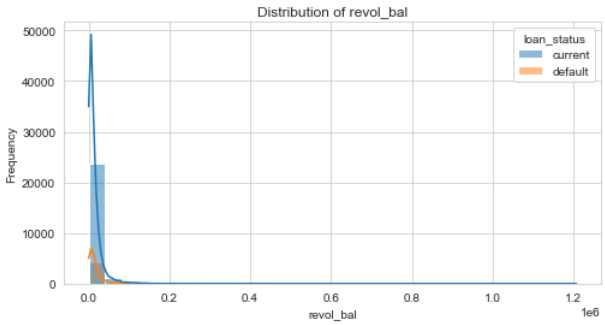

Distribution of fico_range_low
loan_status
current
default
Frequency
fico_range_low

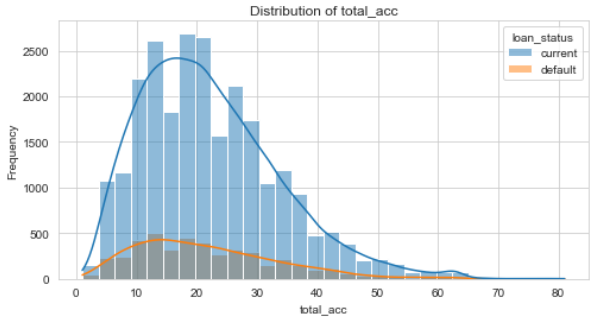

Distribution of open_acc
loan_status
current
default
Frequency
open_acc

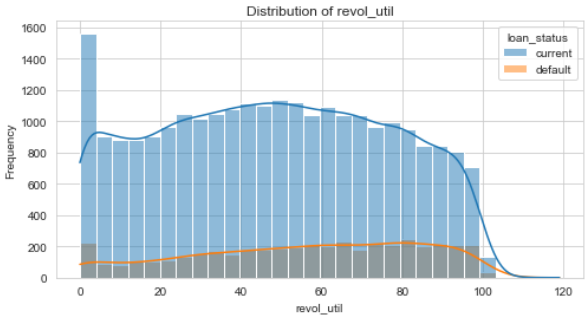

Distribution of fico_range_high
loan_status
current
default
Frequency
fico_range_high

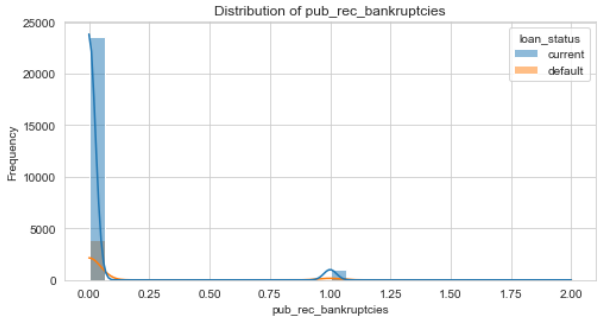

Distribution of pub_rec
loan_status
current
default
Frequency
pub_rec

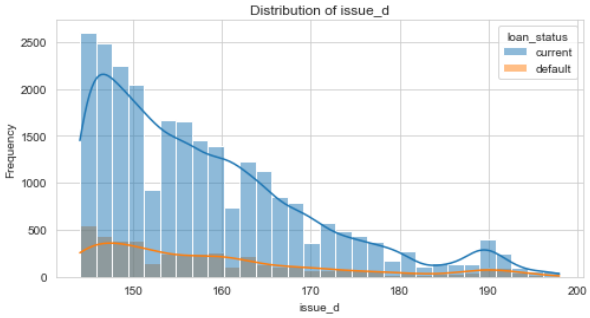

Distribution of inq_last_6mths
loan_status
current
default
Frequency
inq_last_6mths

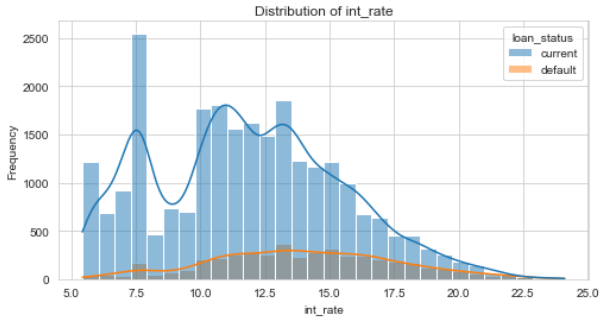

Distribution of revol_bal
loan_status
current
default
Frequency
revol_bal
1e6

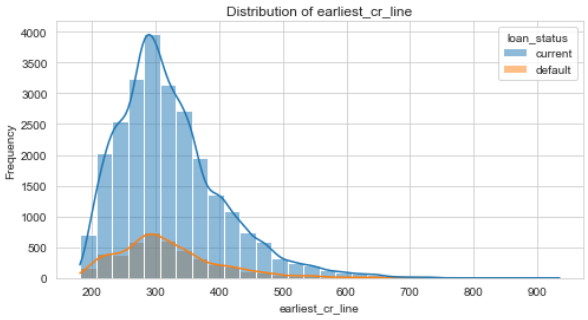


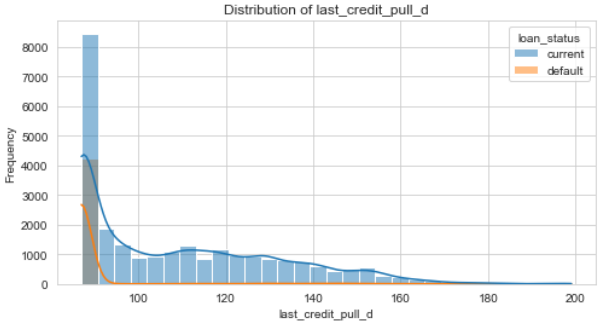




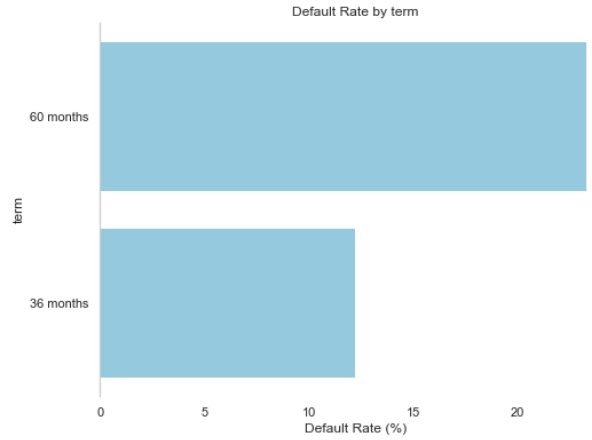


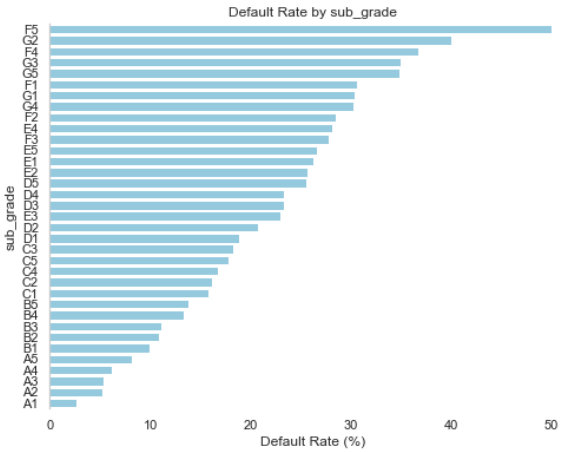


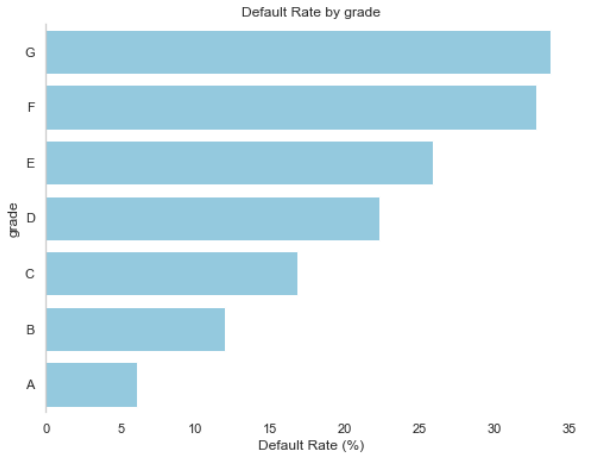


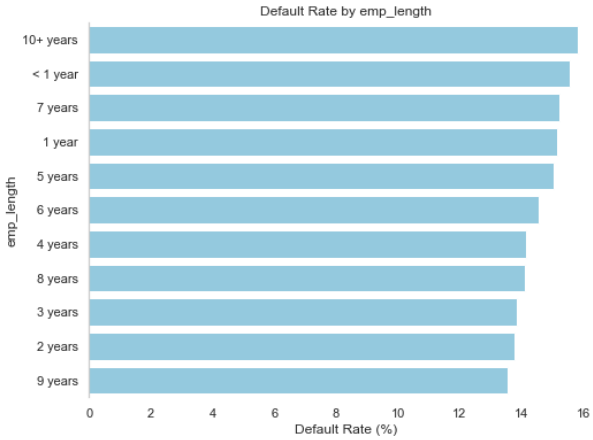


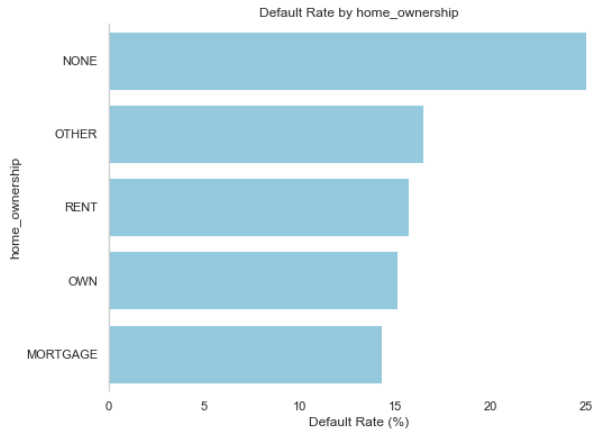
EDA on Categorical Features

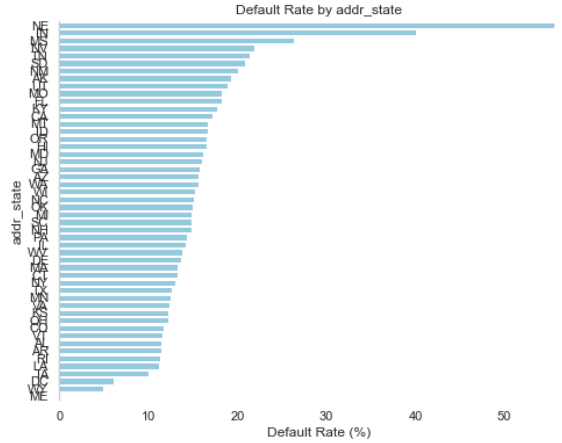

Default Rate by term
60 months
36 months
term
0
5
10
15
20
Default Rate (%)

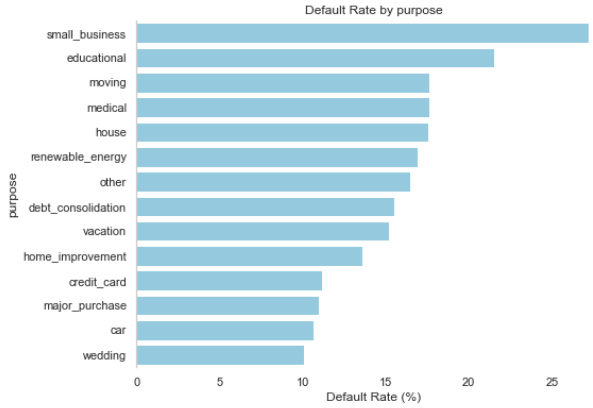

Default Rate by sub_grade
sub_grade
0
10
20
30
40
50
Default Rate (%)

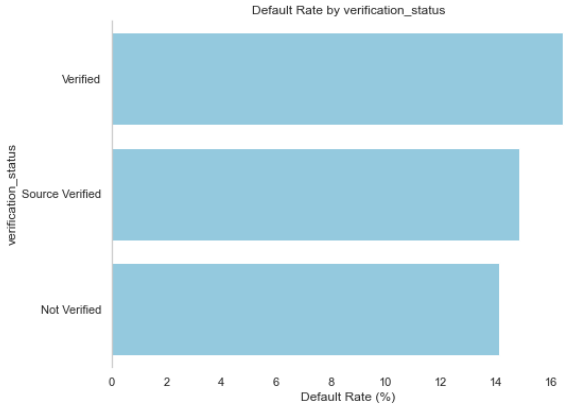

Default Rate by grade
G
F
E
D
C
B
A
grade
0
5
10
15
20
25
30
35
Default Rate (%)


Default Rate by emp_length
10+ years
< 1 year
7 years
1 year
5 years
6 years
4 years
8 years
3 years
2 years
9 years
emp_length
0
2
4
6
8
10
12
14
16
Default Rate (%)

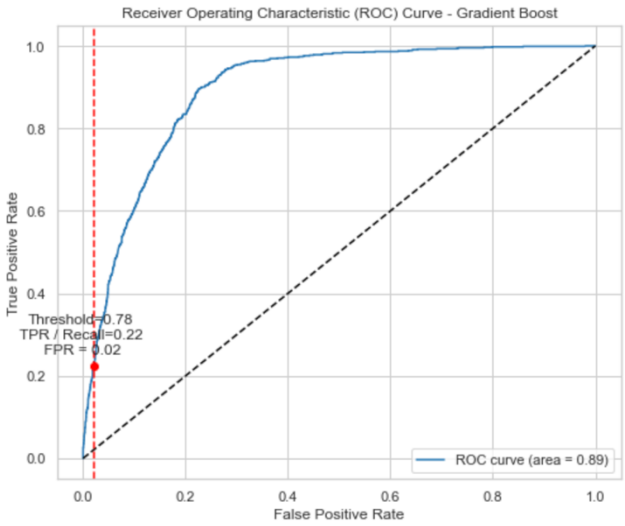

Default Rate by home_ownership
NONE
OTHER
RENT
OWN
MORTGAGE
home_ownership
0
5
10
15
20
25
Default Rate (%)

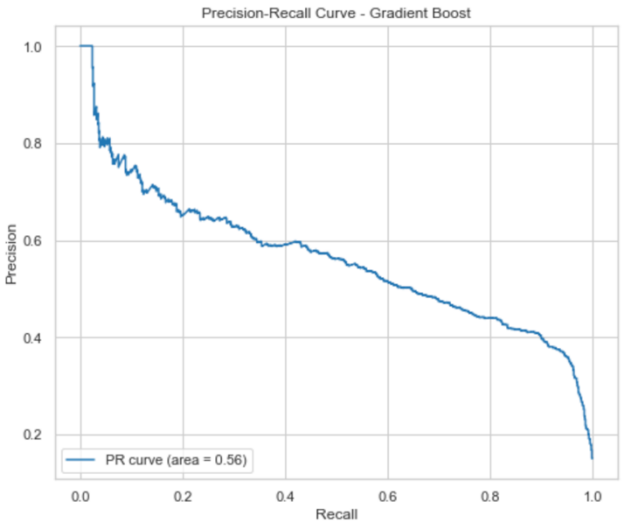

Default Rate by addr_state
addr_state
0
10
20
30
40
50
Default Rate (%)





ROC Graph and Precision-Recall Curve at 2% FPR





Operating Tables

Operating Table – Logistic Regression:

	Target FPR (%)	Expected TPR	Threshold
0	1.0	0.1078	0.7584
1	2.0	0.1703	0.7170
2	3.0	0.2418	0.6796
3	4.0	0.2928	0.6532
4	5.0	0.3314	0.6312
5	6.0	0.3678	0.6118
6	7.0	0.4109	0.5944
7	8.0	0.4415	0.5792
8	9.0	0.4699	0.5640
9	10.0	0.4983	0.5507

Operating Table – Random Forest:

	Target FPR (%)	Expected TPR	Threshold
0	1.0	0.1180	0.6371
1	2.0	0.1986	0.5980
2	3.0	0.2350	0.5810
3	4.0	0.3087	0.5595
4	5.0	0.3553	0.5445
5	6.0	0.3995	0.5282
6	7.0	0.4415	0.5114
7	8.0	0.4677	0.4982
8	9.0	0.5301	0.4836
9	10.0	0.5607	0.4715

Operating Table - Gradient Boost:

	Target FPR (%)	Expected TPR	Threshold
0	1.0	0.1419	0.8130
1	2.0	0.2236	0.7798
2	3.0	0.2951	0.7544
3	4.0	0.3451	0.7362
4	5.0	0.4234	0.7132
5	6.0	0.4654	0.6937
6	7.0	0.5131	0.6732
7	8.0	0.5471	0.6569
8	9.0	0.5800	0.6398
9	10.0	0.6061	0.6248

Operating Table - Neural Network:

	Target FPR (%)	Expected TPR	Threshold
0	1.0	0.1271	0.9355
1	2.0	0.2111	0.9048
2	3.0	0.2622	0.8766
3	4.0	0.3110	0.8520
4	5.0	0.3587	0.8251
5	6.0	0.4154	0.7938
6	7.0	0.4484	0.7685
7	8.0	0.4915	0.7389
8	9.0	0.5335	0.7100
9	10.0	0.5687	0.6829

Operating Table - Stacking Classifier:

	Target FPR (%)	Expected TPR	Threshold
0	1.0	0.1419	0.8354
1	2.0	0.2247	0.8087
2	3.0	0.3019	0.7796
3	4.0	0.3734	0.7560
4	5.0	0.4313	0.7311
5	6.0	0.4711	0.7094
6	7.0	0.4949	0.6886
7	8.0	0.5471	0.6611
8	9.0	0.5880	0.6375
9	10.0	0.6186	0.6111