Assignment 6 - Fannie Mae

Business Case: Fannie Mae is a federally sanctioned corporation that promotes property ownership by buying up privately issued mortgages. Fannie Mae among others received its share of criticism after the mortgage crisis of 2007. One of the many challenges was to identify the defaulters for loan repayments. A model needs to be devised so that the loan can be given to the qualified borrowers who were more likely to make the loan repayment on time. Here we are trying different models on the given data set for Q1 2007 (consisting of 211088 entries) to evaluate as to which customers are more likely to default on loan repayment. This will help Fannie Mae to identify such customers and make informed decisions to reduce potential credit risks. This will help Fannie Mae to gain stability in the market and to continue and provide property ownership by buying up privately issued mortgages.

Payoff matrix:

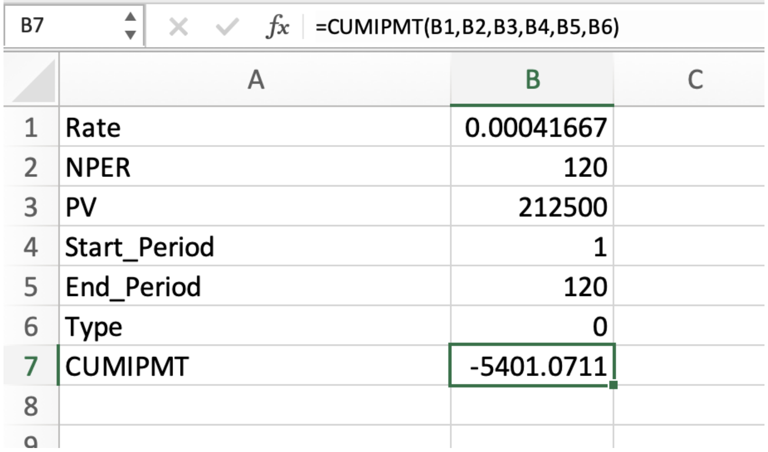
Average house price : $250,000
Average loan amount: 85% of the house price = 250,000 x 0.85=$212,500
Average loan repayment period:10 years(120 months)
Average interest rate on loan : 6%
Average interest earned by Fannie Mae: 0.5%

Case where a customer is a defaulter:
Assumptions:
The house is sold at 90% of the original cost: $225,000
It took Fannie Mae 2 years to sell the property.
Fannie Mae had to spend $20,000 for the renovation of the house before selling.
Some legal cost of roughly $10,000 were involved against the defaulter.

True Positive:
The model predicts that the customer will be a defaulter and hence Fannie Mae does not provide Loan. No loan is issued hence there is no profit.
TP: $0

True Negative: The model predicts that the customer will not be a defaulter and hence Fannie Mae does provide a loan. As loan is issued hence there is profit generated.

TN: $5401.0711, approx. $5401



B7 =CUMIPMT(B1,B2,B3,B4,B5,B6)

	A	B	C
1	Rate	0.00041667	
2	NPER	120	
3	PV	212500	
4	Start_Period	1	
5	End_Period	120	
6	Type	0	
7	CUMIPMT	-5401.0711	
8			
9			

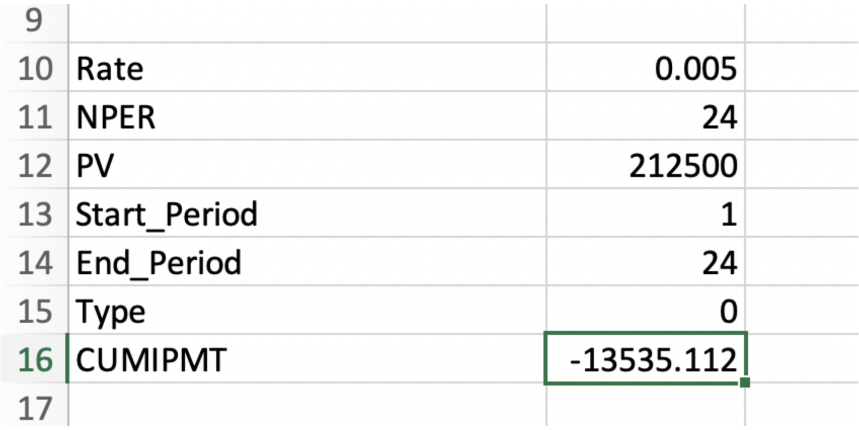
False Positive:The model incorrectly identifies as a potential customer as defaulter and therefore no loan is sanctioned by Fannie Mae and therefore no profit is generated. This could be considered as an opportunity cost.

FP: $0

False Negative: The model identifies a customer who is likely to default as a non-defaulter and hence Fannie Mae provides the loan and must bear the losses for the same. Assuming the customer did no repayment and defaulted from the 1st month itself.

$225,000(price at which the house is sold) -$212,500(loan amount, that is 85% of the house price)-$13535.112(interest rate for 2 years)- $20,000(renovation cost)-$10,000 (legal cost)

FN: -31,035.112 , approx. $-31,035



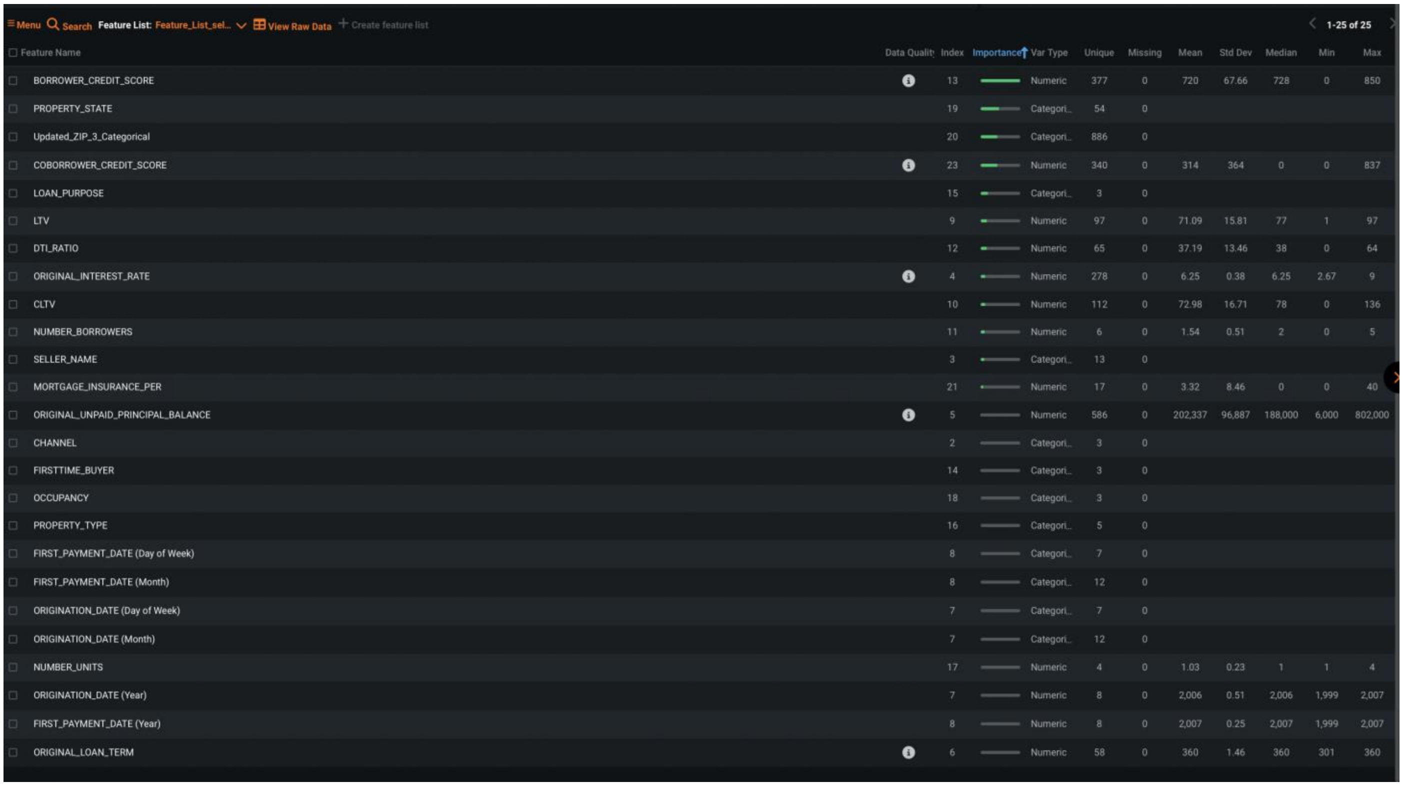
9			
10	Rate	0.005	
11	NPER	24	
12	PV	212500	
13	Start_Period	1	
14	End_Period	24	
15	Type	0	
16	CUMIPMT	-13535.112	
17			

	Predicted 0	Predicted 1
Actual 0	5401	0
Actual 1	-31,035	0

Question 1:

Data pre-processing :

1. Excluded the variable due to high cardinality : Loan ID
2. Excluded the variable product_type as it only had 1 value for all the data entries and hence would not make any difference in model prediction.
3. Data Robot converted FIRST_PAYMENT_DATE into FIRST_PAYMENT_DATE (Day of Week), FIRST_PAYMENT_DATE (Month), FIRST_PAYMENT_DATE (Year) , and FIRST_PAYMENT_DATE (Day of Month) essentially extracting all the important information from the given date and hence these new features are used and FIRST_PAYMENT_DATE is excluded.
4. Data Robot converted ORIGINATION_DATE into ORIGINATION_DATE (Day of Week), ORIGINATION_DATE (Month), ORIGINATION_DATE (Year) and ORIGINATION_DATE (Day of Month) essentially extracting all the important information from the given date and hence these new features are used and ORIGINATION_DATE is excluded.
5. FIRST_PAYMENT_DATE (Day of Month) and ORIGINATION_DATE (Day of Month) are excluded from the feature list as it contains only 1 value for the entire dataset.



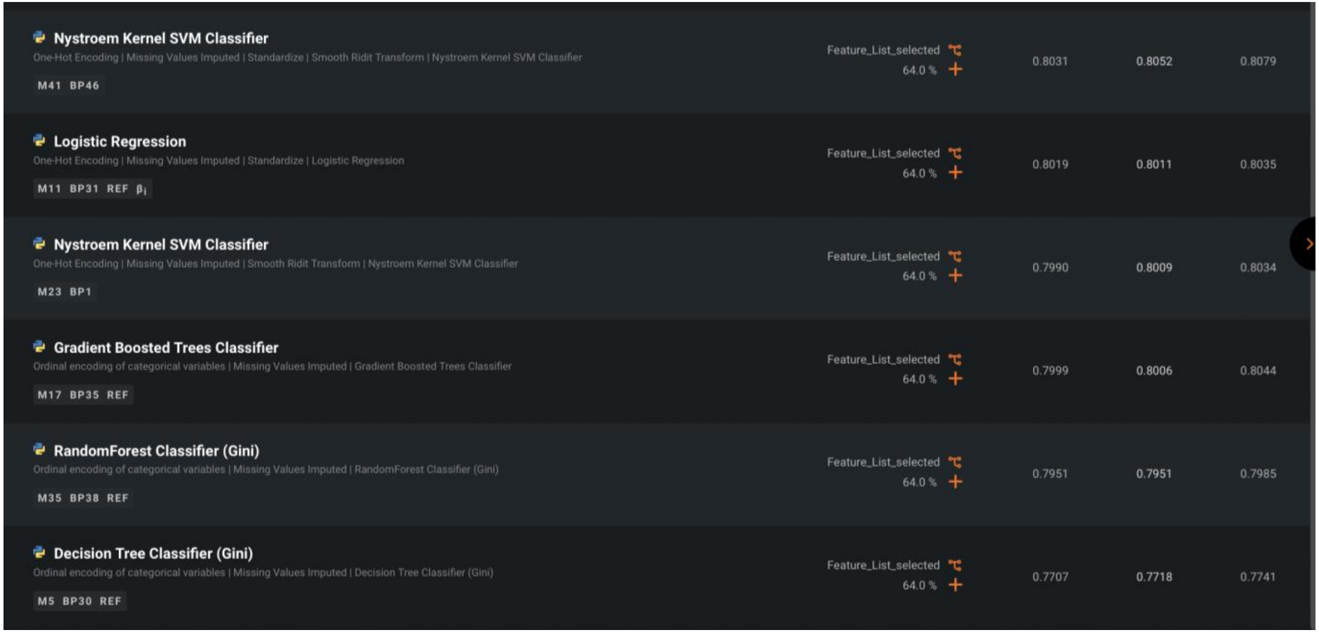
≡ Menu Search Feature List: Feature_List_sel... View Raw Data + Create feature list

1-25 of 25

□ Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
□ BORROWER_CREDIT_SCORE	ⓘ	13		Numeric	377	0	720	67.66	728	0	850
□ PROPERTY_STATE		19		Categori...	54	0					
□ Updated_ZIP_3_Categorical		20		Categori...	886	0					
□ COBORROWER_CREDIT_SCORE	ⓘ	23		Numeric	340	0	314	364	0	0	837
□ LOAN_PURPOSE		15		Categori...	3	0					
□ LTV		9		Numeric	97	0	71.09	15.81	77	1	97
□ DTI_RATIO		12		Numeric	65	0	37.19	13.46	38	0	64
□ ORIGINAL_INTEREST_RATE	ⓘ	4		Numeric	278	0	6.25	0.38	6.25	2.67	9
□ CLTV		10		Numeric	112	0	72.98	16.71	78	0	136
□ NUMBER_BORROWERS		11		Numeric	6	0	1.54	0.51	2	0	5
□ SELLER_NAME		3		Categori...	13	0					
□ MORTGAGE_INSURANCE_PER		21		Numeric	17	0	3.32	8.46	0	0	40
□ ORIGINAL_UNPAID_PRINCIPAL_BALANCE	ⓘ	5		Numeric	586	0	202,337	96,887	188,000	6,000	802,000
□ CHANNEL		2		Categori...	3	0					
□ FIRSTTIME_BUYER		14		Categori...	3	0					
□ OCCUPANCY		18		Categori...	3	0					
□ PROPERTY_TYPE		16		Categori...	5	0					
□ FIRST_PAYMENT_DATE (Day of Week)		8		Categori...	7	0					
□ FIRST_PAYMENT_DATE (Month)		8		Categori...	12	0					
□ ORIGINATION_DATE (Day of Week)		7		Categori...	7	0					
□ ORIGINATION_DATE (Month)		7		Categori...	12	0					
□ NUMBER_UNITS		17		Numeric	4	0	1.03	0.23	1	1	4
□ ORIGINATION_DATE (Year)		7		Numeric	8	0	2,006	0.51	2,006	1,999	2,007
□ FIRST_PAYMENT_DATE (Year)		8		Numeric	8	0	2,007	0.25	2,007	1,999	2,007
□ ORIGINAL_LOAN_TERM	ⓘ	6		Numeric	58	0	360	1.46	360	301	360

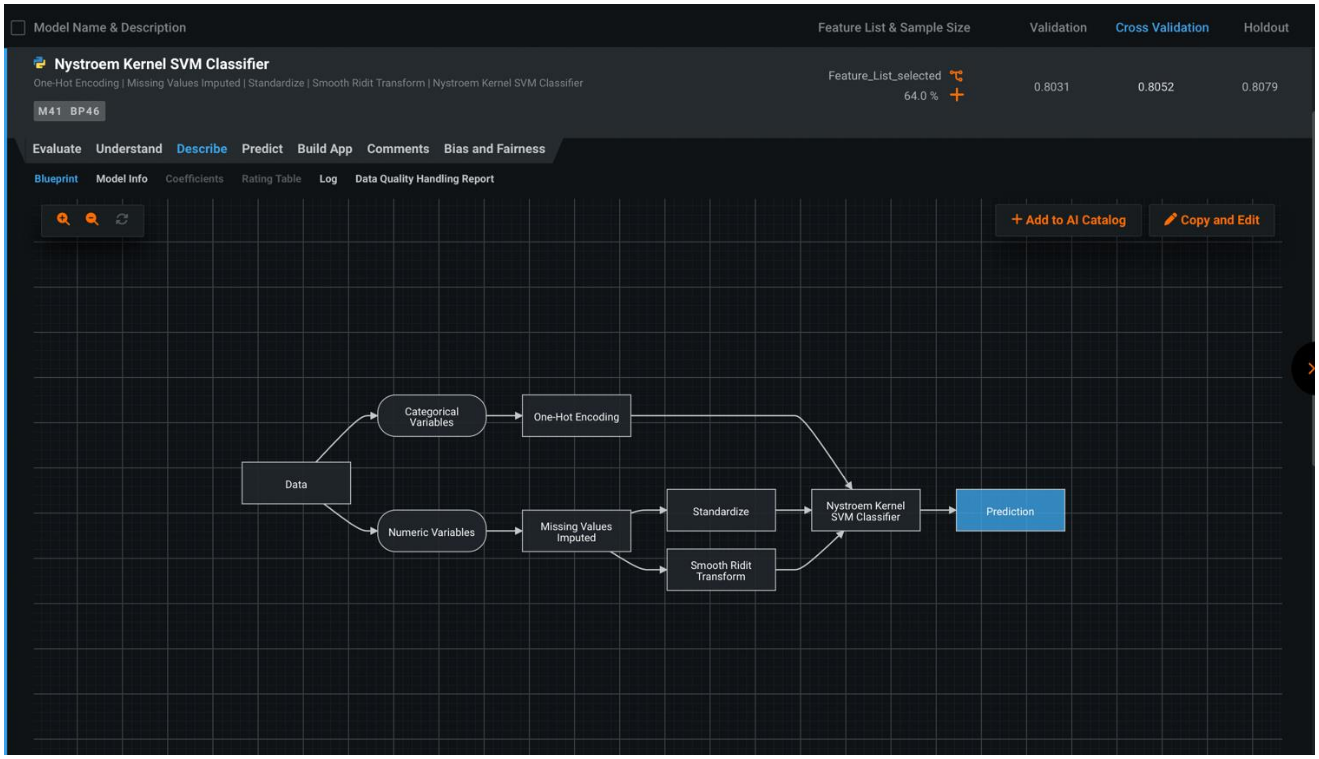
Question 2:

The following models were run on the given data:

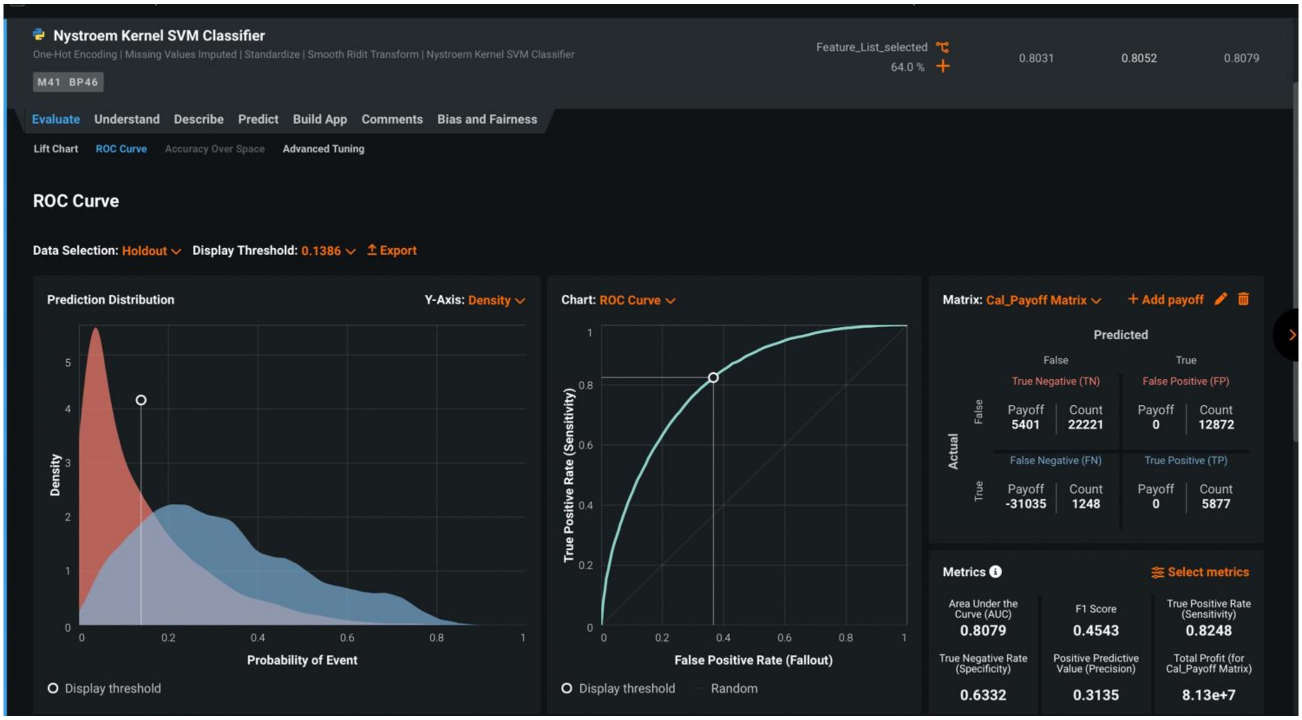


Nystroem Kernel SVM Classifier One-Hot Encoding | Missing Values Imputed | Standardize | Smooth Ridit Transform | Nystroem Kernel SVM Classifier M41 BP46	Feature_List_selected 64.0 % +	0.8031	0.8052	0.8079
Logistic Regression One-Hot Encoding | Missing Values Imputed | Standardize | Logistic Regression M11 BP31 REF β	Feature_List_selected 64.0 % +	0.8019	0.8011	0.8035
Nystroem Kernel SVM Classifier One-Hot Encoding | Missing Values Imputed | Smooth Ridit Transform | Nystroem Kernel SVM Classifier M23 BP1	Feature_List_selected 64.0 % +	0.7990	0.8009	0.8034
Gradient Boosted Trees Classifier Ordinal encoding of categorical variables | Missing Values Imputed | Gradient Boosted Trees Classifier M17 BP35 REF	Feature_List_selected 64.0 % +	0.7999	0.8006	0.8044
RandomForest Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Classifier (Gini) M35 BP38 REF	Feature_List_selected 64.0 % +	0.7951	0.7951	0.7985
Decision Tree Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | Decision Tree Classifier (Gini) M5 BP30 REF	Feature_List_selected 64.0 % +	0.7707	0.7718	0.7741

1.Nystroem Kernel SVM Classifier:



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Nystroem Kernel SVM Classifier One-Hot Encoding | Missing Values Imputed | Standardize | Smooth Ridit Transform | Nystroem Kernel SVM Classifier M41 BP46	Feature_List_selected 64.0 %	0.8031	0.8052	0.8079



	Feature List & Sample Size	Validation	Cross Validation	Holdout
Nystroem Kernel SVM Classifier One-Hot Encoding | Missing Values Imputed | Standardize | Smooth Ridit Transform | Nystroem Kernel SVM Classifier M41 BP46	Feature_List_selected 64.0 %	0.8031	0.8052	0.8079

ROC Curve

Data Selection: Holdout Display Threshold: 0.1386 Export

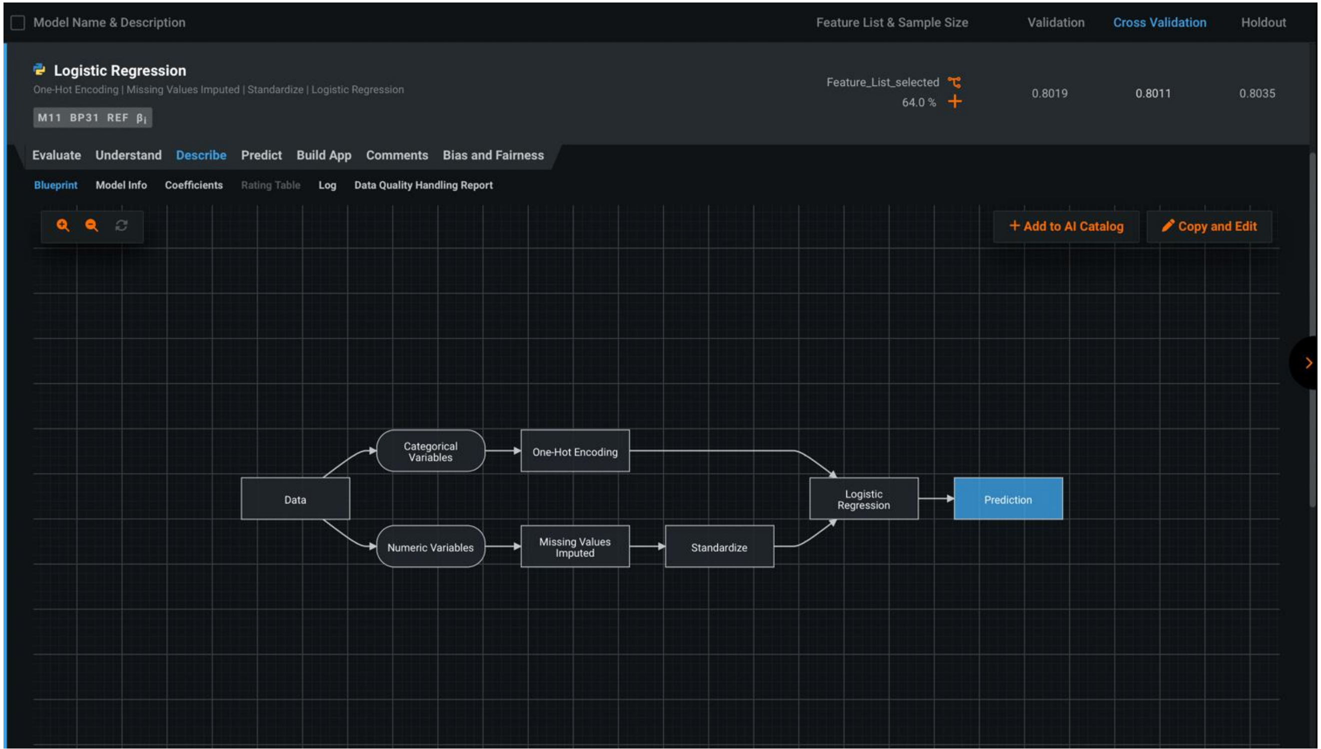
Matrix: Cal_Payoff Matrix

		Predicted False		Predicted True	
		True Negative (TN)		False Positive (FP)	
		Payoff	Count	Payoff	Count
Actual	False	5401	22221	0	12872
		False Negative (FN)		True Positive (TP)	
		Payoff	Count	Payoff	Count
Actual	True	-31035	1248	0	5877

Metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.8079	0.4543	0.8248
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Cal_Payoff Matrix)**
0.6332	0.3135	8.13e+7

2. Logistic Regression:

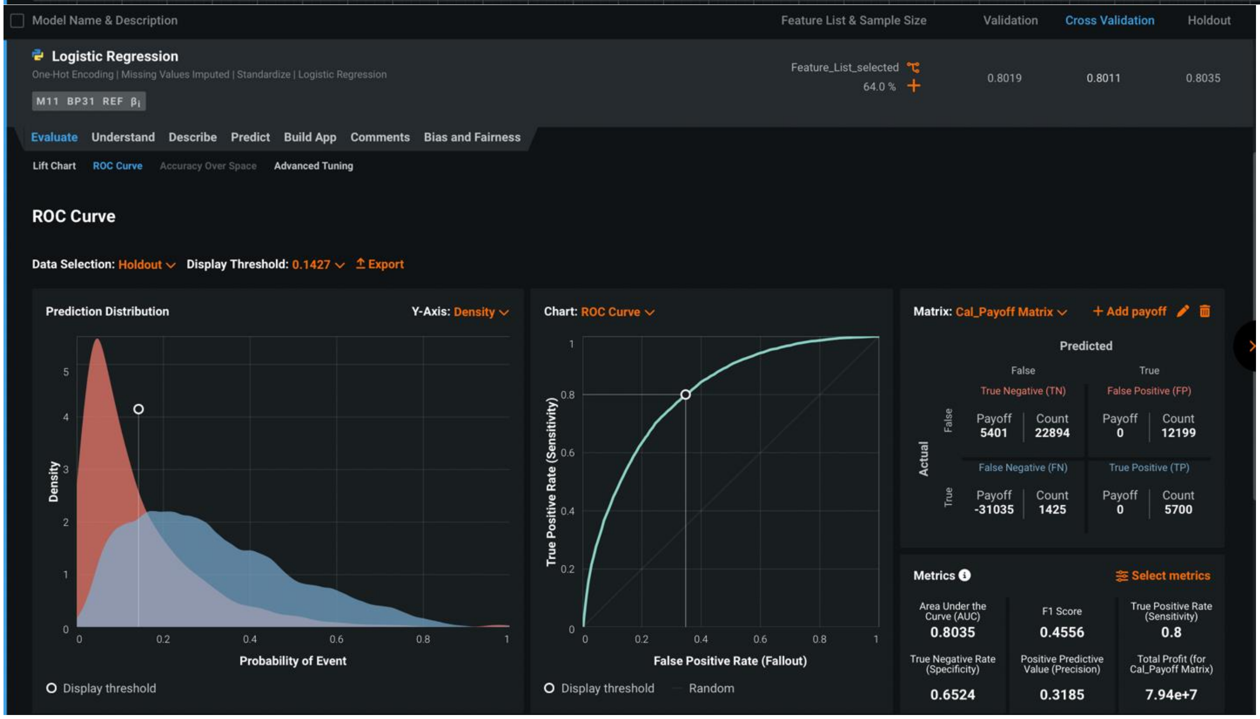


Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Logistic Regression One-Hot Encoding | Missing Values Imputed | Standardize | Logistic Regression M11 BP31 REF βᵢ	Feature_List_selected 64.0 %	0.8019	0.8011	0.8035

Evaluate Understand **Describe** Predict Build App Comments Bias and Fairness

Blueprint Model Info Coefficients Rating Table Log Data Quality Handling Report

+ Add to AI Catalog   Copy and Edit



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Logistic Regression One-Hot Encoding | Missing Values Imputed | Standardize | Logistic Regression M11 BP31 REF βᵢ	Feature_List_selected 64.0 %	0.8019	0.8011	0.8035

Evaluate Understand Describe Predict Build App Comments Bias and Fairness

Lift Chart **ROC Curve** Accuracy Over Space Advanced Tuning

ROC Curve

Data Selection: Holdout   Display Threshold: 0.1427   Export

Prediction Distribution   Y-Axis: Density

Chart: ROC Curve

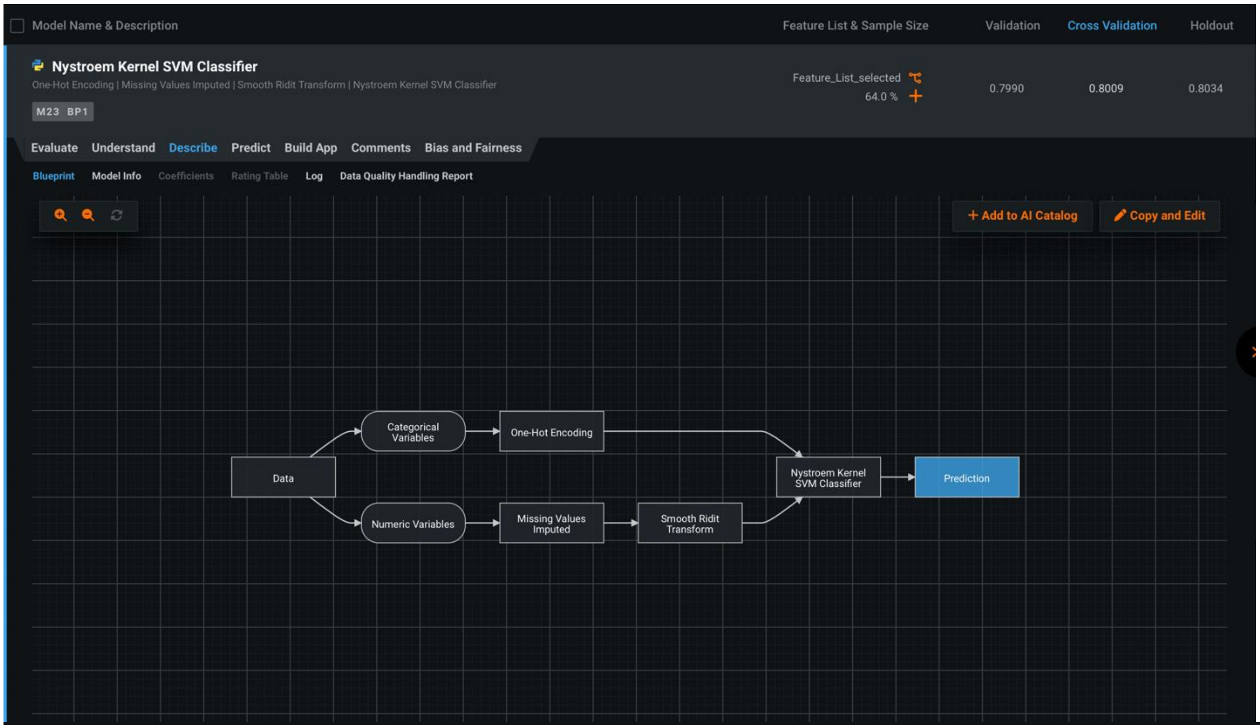
Matrix: Cal_Payoff Matrix   + Add payoff

		Predicted False	Predicted True
Actual	False	True Negative (TN) Payoff 5401 | Count 22894	False Positive (FP) Payoff 0 | Count 12199
Actual	True	False Negative (FN) Payoff -31035 | Count 1425	True Positive (TP) Payoff 0 | Count 5700

Metrics   Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.8035	0.4556	0.8
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Cal_Payoff Matrix)**
0.6524	0.3185	7.94e+7

3.Nystroem Kernel SVM Classifier



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Nystroem Kernel SVM Classifier One-Hot Encoding | Missing Values Imputed | Smooth Ridit Transform | Nystroem Kernel SVM Classifier M23 BP1	Feature_List_selected 64.0 %	0.7990	0.8009	0.8034

Evaluate Understand Describe Predict Build App Comments Bias and Fairness

Blueprint Model Info Coefficients Rating Table Log Data Quality Handling Report

+ Add to AI Catalog ✎ Copy and Edit

Data

Categorical Variables

One-Hot Encoding

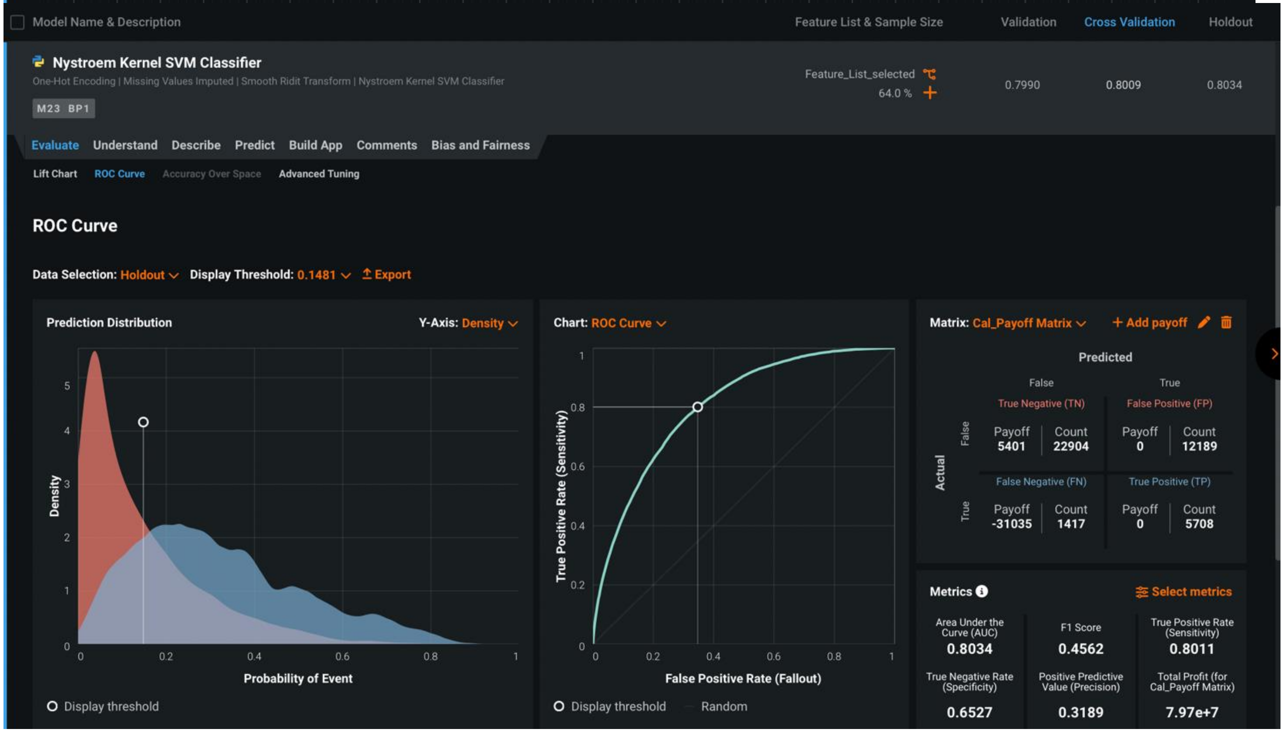
Numeric Variables

Missing Values Imputed

Smooth Ridit Transform

Nystroem Kernel SVM Classifier

Prediction



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Nystroem Kernel SVM Classifier One-Hot Encoding | Missing Values Imputed | Smooth Ridit Transform | Nystroem Kernel SVM Classifier M23 BP1	Feature_List_selected 64.0 %	0.7990	0.8009	0.8034

Evaluate Understand Describe Predict Build App Comments Bias and Fairness

Lift Chart ROC Curve Accuracy Over Space Advanced Tuning

ROC Curve

Data Selection: Holdout **Display Threshold:** 0.1481 Export

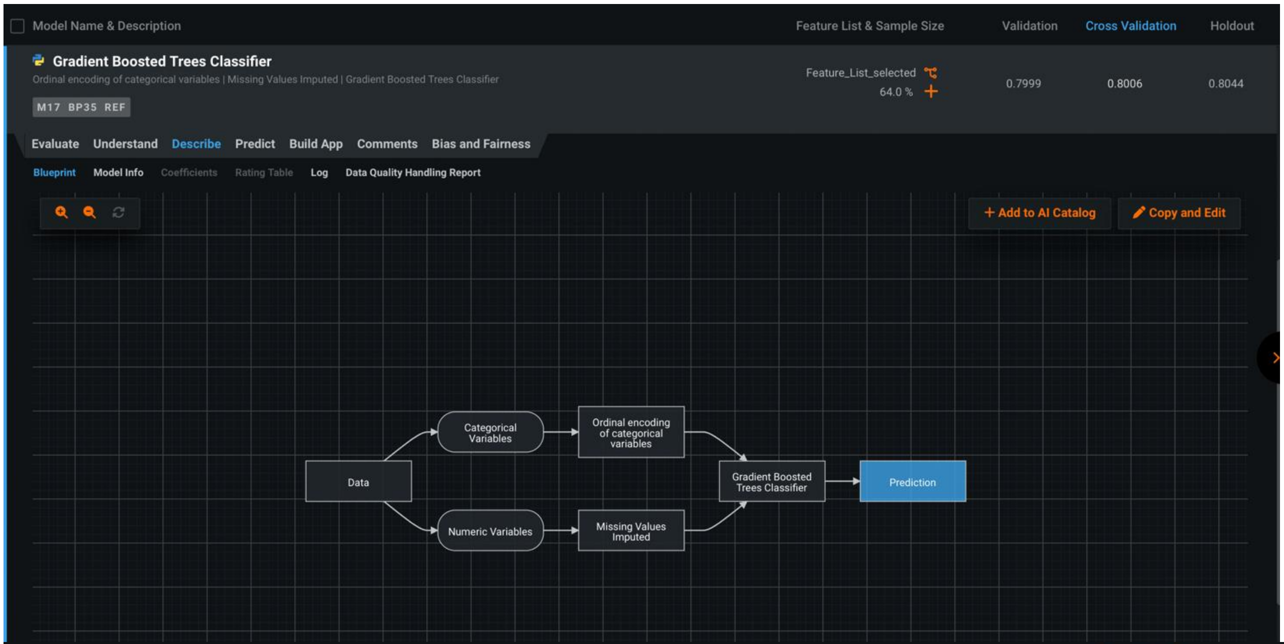
Matrix: Cal_Payoff Matrix + Add payoff

		Predicted False		Predicted True	
		True Negative (TN)		False Positive (FP)	
Actual	False	Payoff 5401	Count 22904	Payoff 0	Count 12189
		False Negative (FN)		True Positive (TP)	
Actual	True	Payoff -31035	Count 1417	Payoff 0	Count 5708

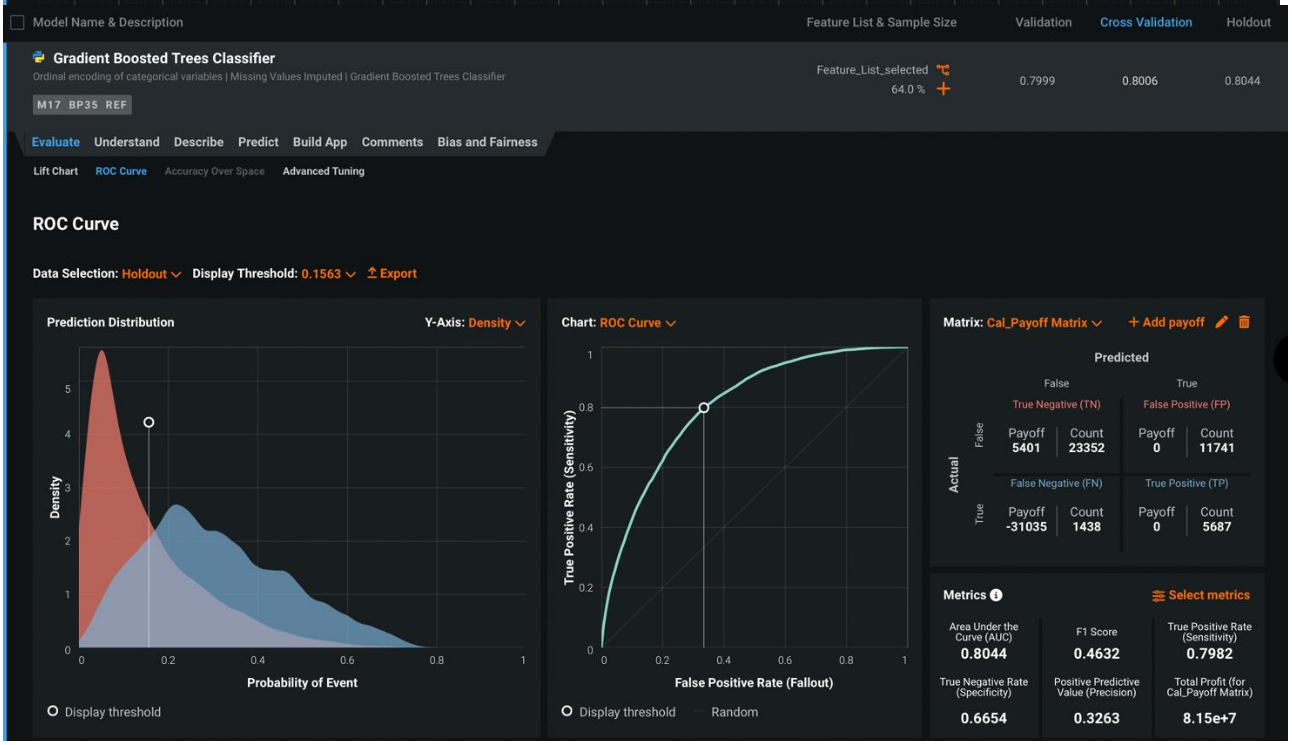
Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.8034	0.4562	0.8011
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Cal_Payoff Matrix)**
0.6527	0.3189	7.97e+7

4. Gradient Boosted Trees Classifier



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Gradient Boosted Trees Classifier Ordinal encoding of categorical variables | Missing Values Imputed | Gradient Boosted Trees Classifier M17 BP35 REF	Feature_List_selected 64.0 %	0.7999	0.8006	0.8044



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Gradient Boosted Trees Classifier Ordinal encoding of categorical variables | Missing Values Imputed | Gradient Boosted Trees Classifier M17 BP35 REF	Feature_List_selected 64.0 %	0.7999	0.8006	0.8044

ROC Curve

Data Selection: Holdout Display Threshold: 0.1563 Export

Matrix: Cal_Payoff Matrix

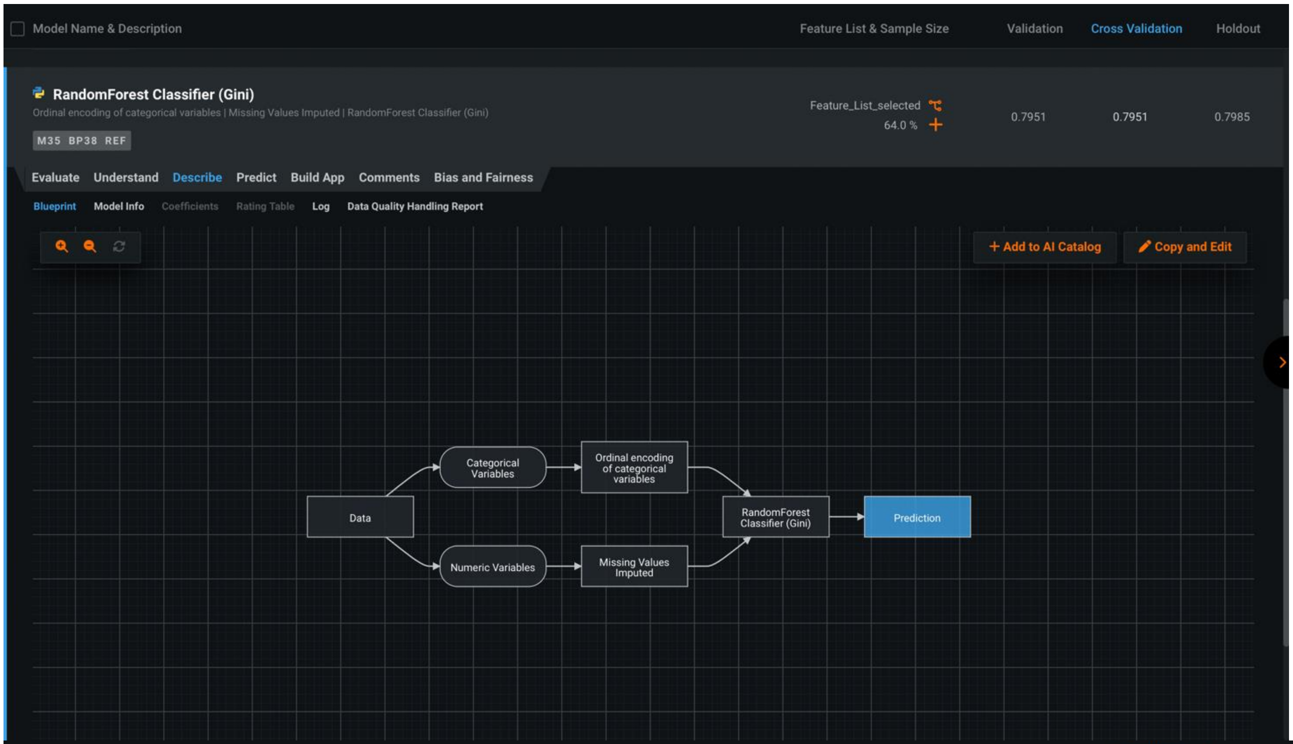
		Predicted: False		Predicted: True	
		True Negative (TN)		False Positive (FP)	
		Payoff	Count	Payoff	Count
Actual	False	5401	23352	0	11741
		False Negative (FN)		True Positive (TP)	
		Payoff	Count	Payoff	Count
Actual	True	-31035	1438	0	5687

Metrics

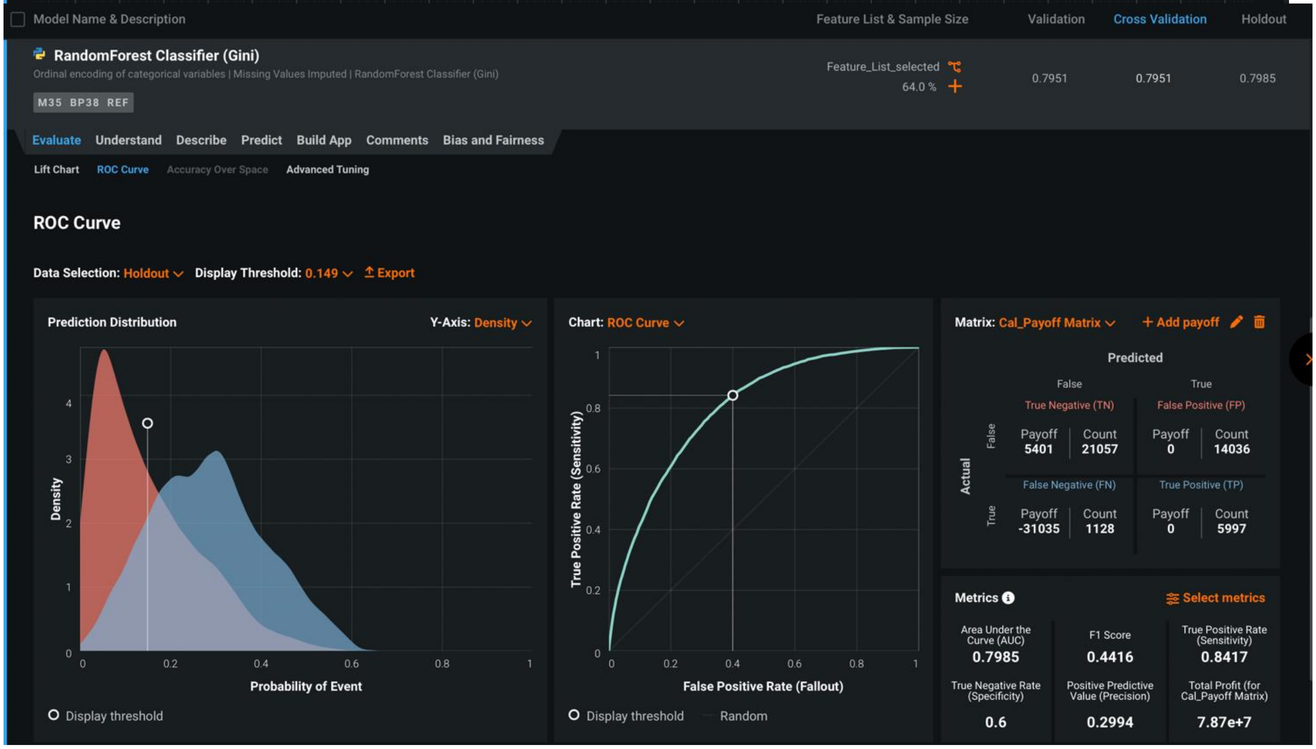
Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.8044	0.4632	0.7982

True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Cal_Payoff Matrix)
0.6654	0.3263	8.15e+7

5. RandomForest Classifier (Gini)



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
RandomForest Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Classifier (Gini) M35 BP38 REF	Feature_List_selected 64.0 %	0.7951	0.7951	0.7985



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
RandomForest Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Classifier (Gini) M35 BP38 REF	Feature_List_selected 64.0 %	0.7951	0.7951	0.7985

ROC Curve

Data Selection: Holdout Display Threshold: 0.149 Export

Matrix: Cal_Payoff Matrix + Add payoff

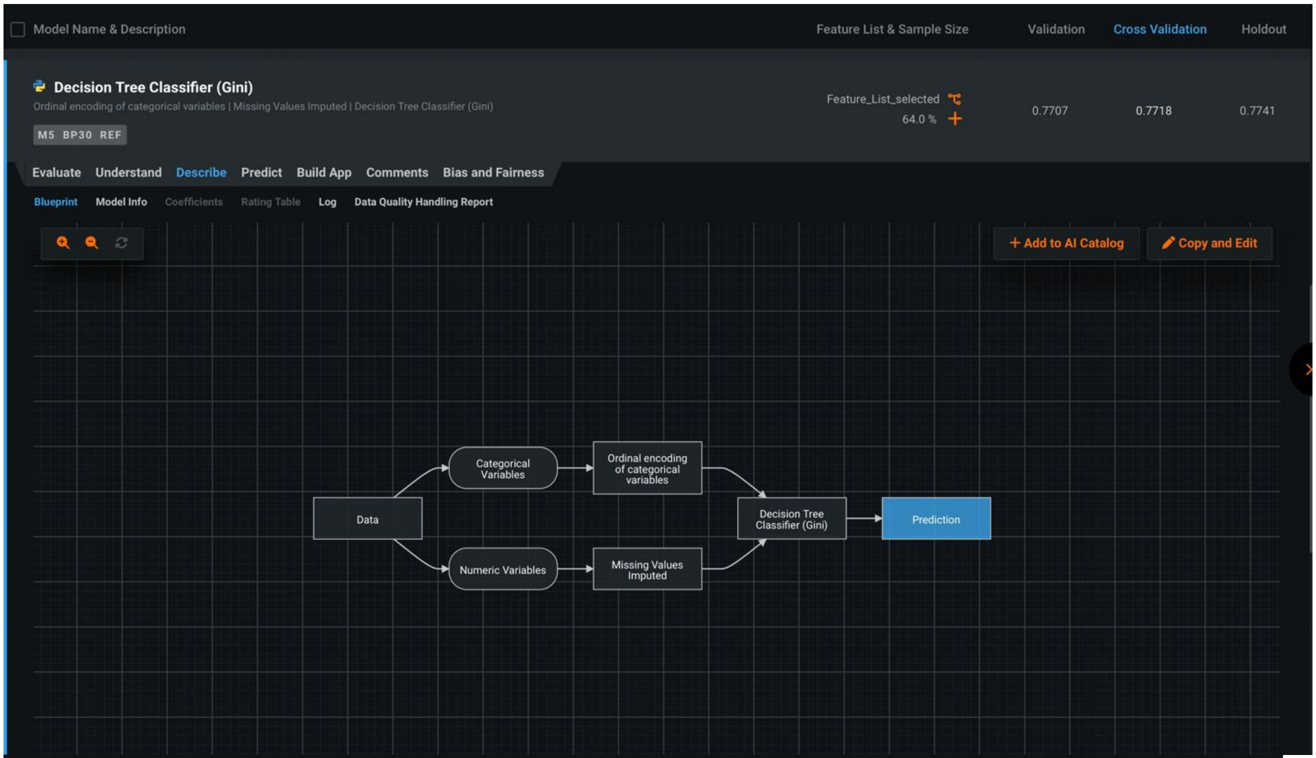
		Predicted False	Predicted True
Actual	False	True Negative (TN) Payoff 5401 Count 21057	False Positive (FP) Payoff 0 Count 14036
	True	False Negative (FN) Payoff -31035 Count 1128	True Positive (TP) Payoff 0 Count 5997

Metrics Select metrics

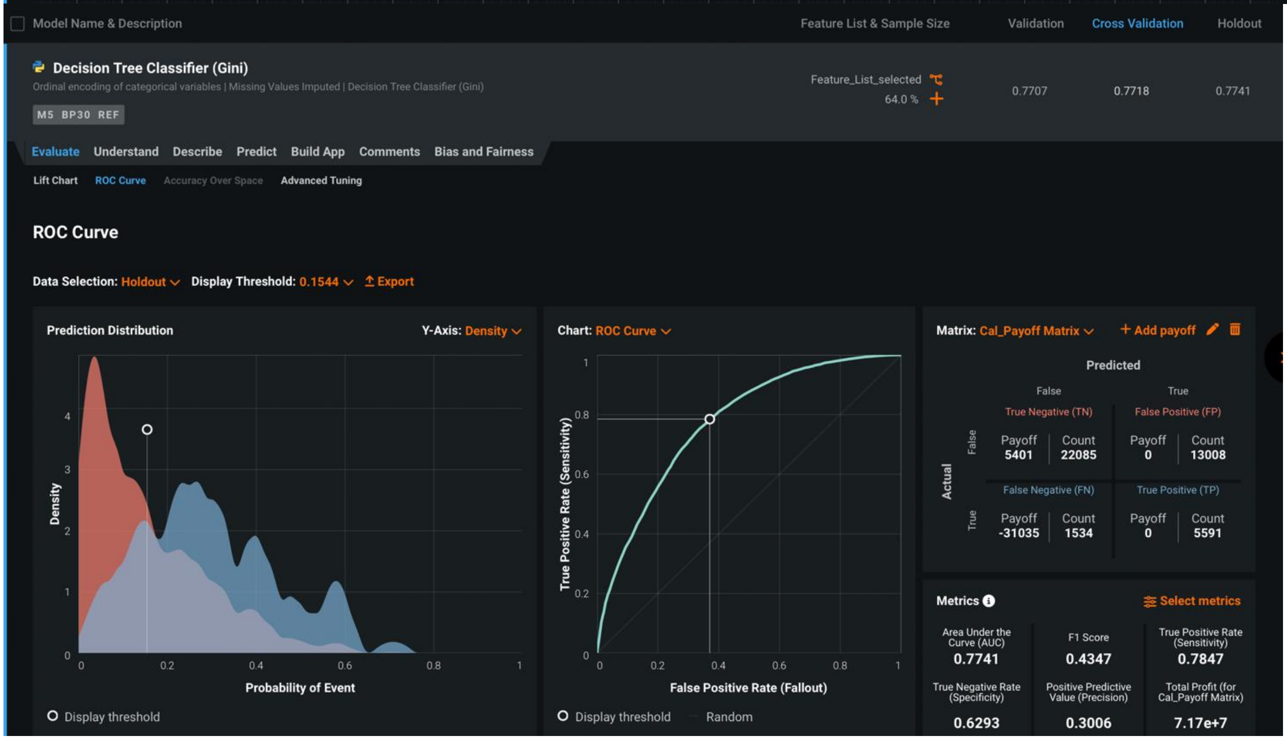
Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.7985	0.4416	0.8417

True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Cal_Payoff Matrix)
0.6	0.2994	7.87e+7

6. Decision Tree Classifier (Gini)



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Decision Tree Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | Decision Tree Classifier (Gini) M5 BP30 REF	Feature_List_selected 64.0 %	0.7707	0.7718	0.7741



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Decision Tree Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | Decision Tree Classifier (Gini) M5 BP30 REF	Feature_List_selected 64.0 %	0.7707	0.7718	0.7741

ROC Curve

Data Selection: Holdout Display Threshold: 0.1544 Export

Matrix: Cal_Payoff Matrix + Add payoff

		Predicted: False	Predicted: True
Actual	False	True Negative (TN) Payoff 5401 | Count 22085	False Positive (FP) Payoff 0 | Count 13008
Actual	True	False Negative (FN) Payoff -31035 | Count 1534	True Positive (TP) Payoff 0 | Count 5591

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.7741	0.4347	0.7847
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Cal_Payoff Matrix)**
0.6293	0.3006	7.17e+7

KNN model is not evaluated here as the data is too complex for its evaluation.

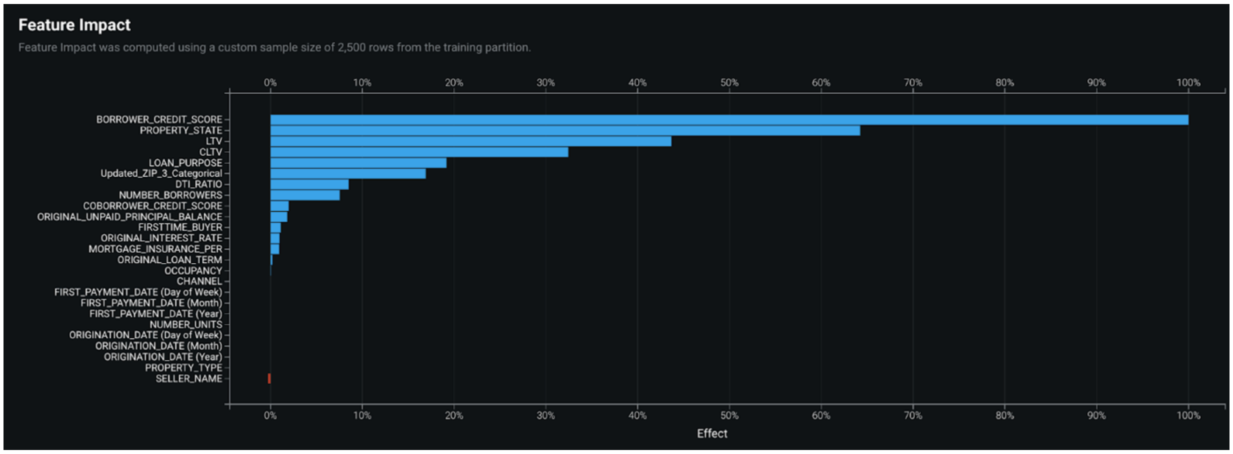
Question 3: Model performance metric

Model	Recall	Precision	Specificity	F1	ROC AUC	Maximum payoff
Nystroem Kernel SVM Classifier(Threshold: 0.1386)	0.8248	0.3135	0.6332	0.4543	0.8079	$81,300,000
Logistic Regression(Threshold: 0.1427)	0.8	0.3185	0.6524	0.4556	0.8035	$79,400,000
Gradient Boosted Trees Classifier(Threshold: 0.1563)	0.7982	0.3263	0.6654	0.4632	0.8044	$81,500,000
RandomForest Classifier (Gini)(Threshold:0.149)	0.8417	0.2994	0.6	0.4416	0.7985	$78,700,000
Decision Tree Classifier (Gini)(Threshold:0.1544)	0.7847	0.3006	0.6293	0.4347	0.7741	$71,700,000

The best performing model is Boosted Tree Classifier. It provides the highest payoff metric of $81,500,000. Best Metric to evaluate the model in our case is the Maximum payoff, as in our case our end goal is to evaluate the profit based on the number of customers who will not default on loan repayment. Maximum payoff metric assigns costs and benefits to different types of correct and incorrect predictions (true positives/true negatives and false positives/false negatives) and help evaluate the required profit/losses based on the given case. As our business case requires us to evaluate the potential customers who will not default in loan repayment and help Fannie Mae gain profit, we need a profit metric to evaluate the same and hence maximum payoff is ideal for this case.

Question 4:

Feature effect:

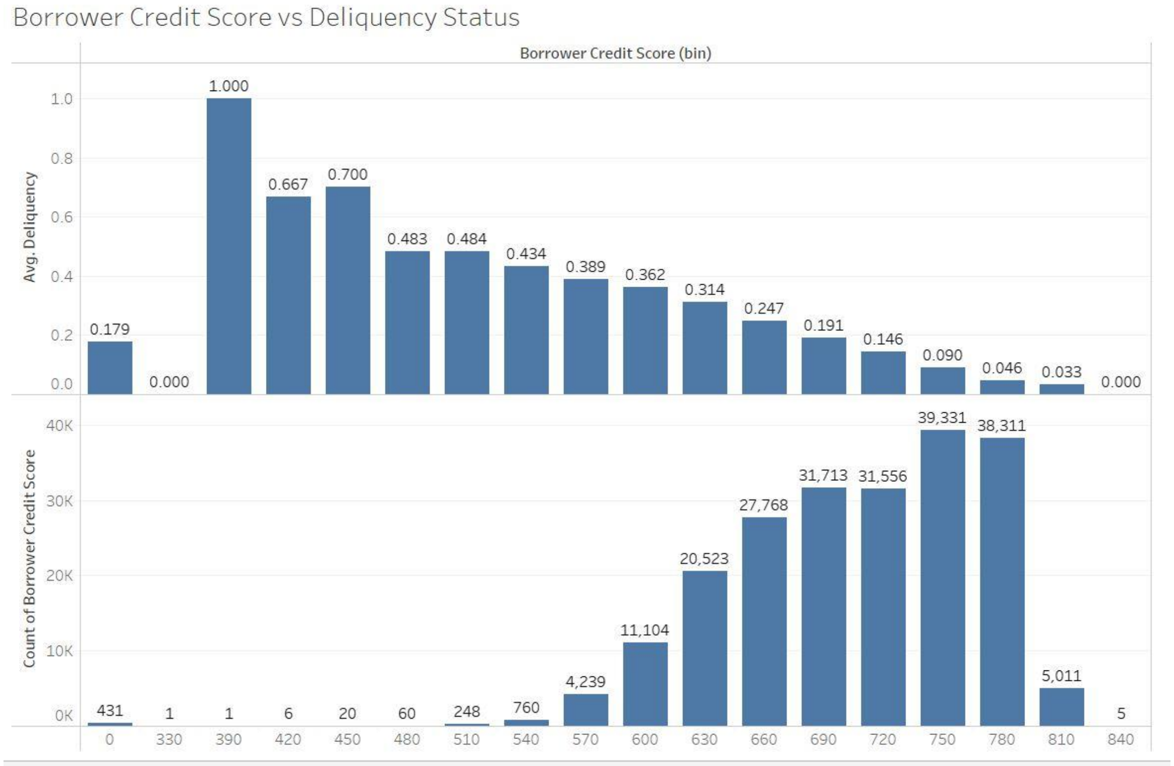


Feature Impact

Feature Impact was computed using a custom sample size of 2,500 rows from the training partition.

Borrower Credit score, Property State , LTV(Loan-to-Value) and CLTV(Combined Loan-to-Value) seem seems to have highest impact among the given features.

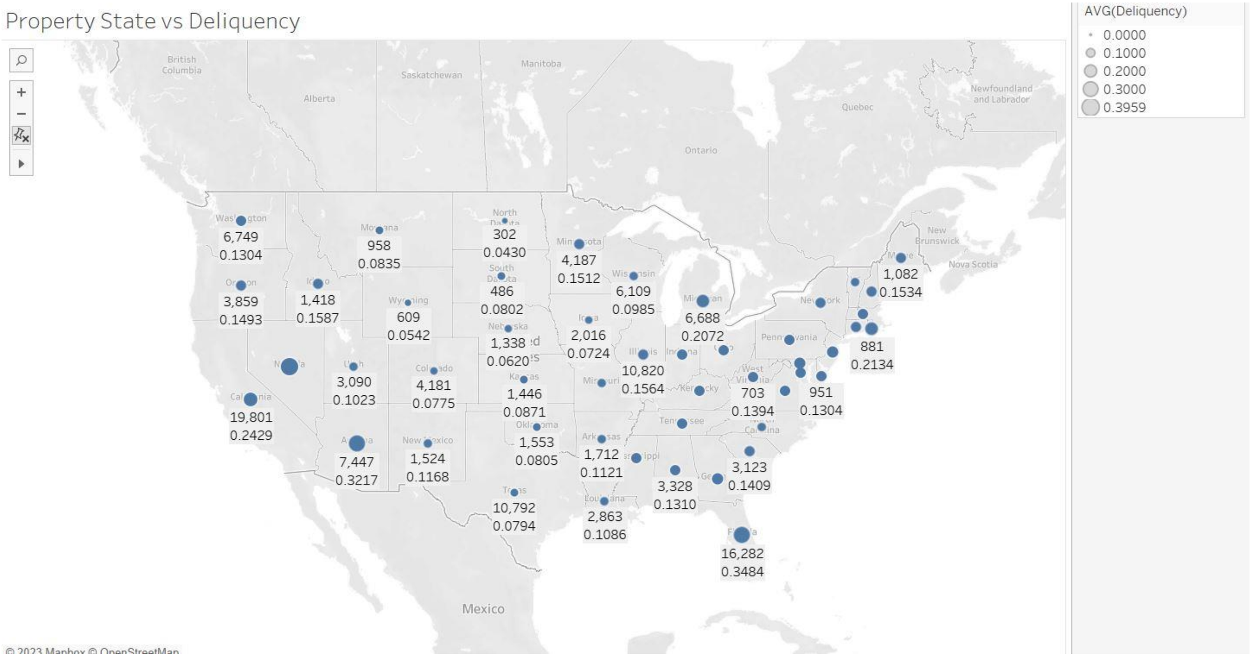
1. Borrower Credit Score



Borrower Credit Score vs Deliquency Status

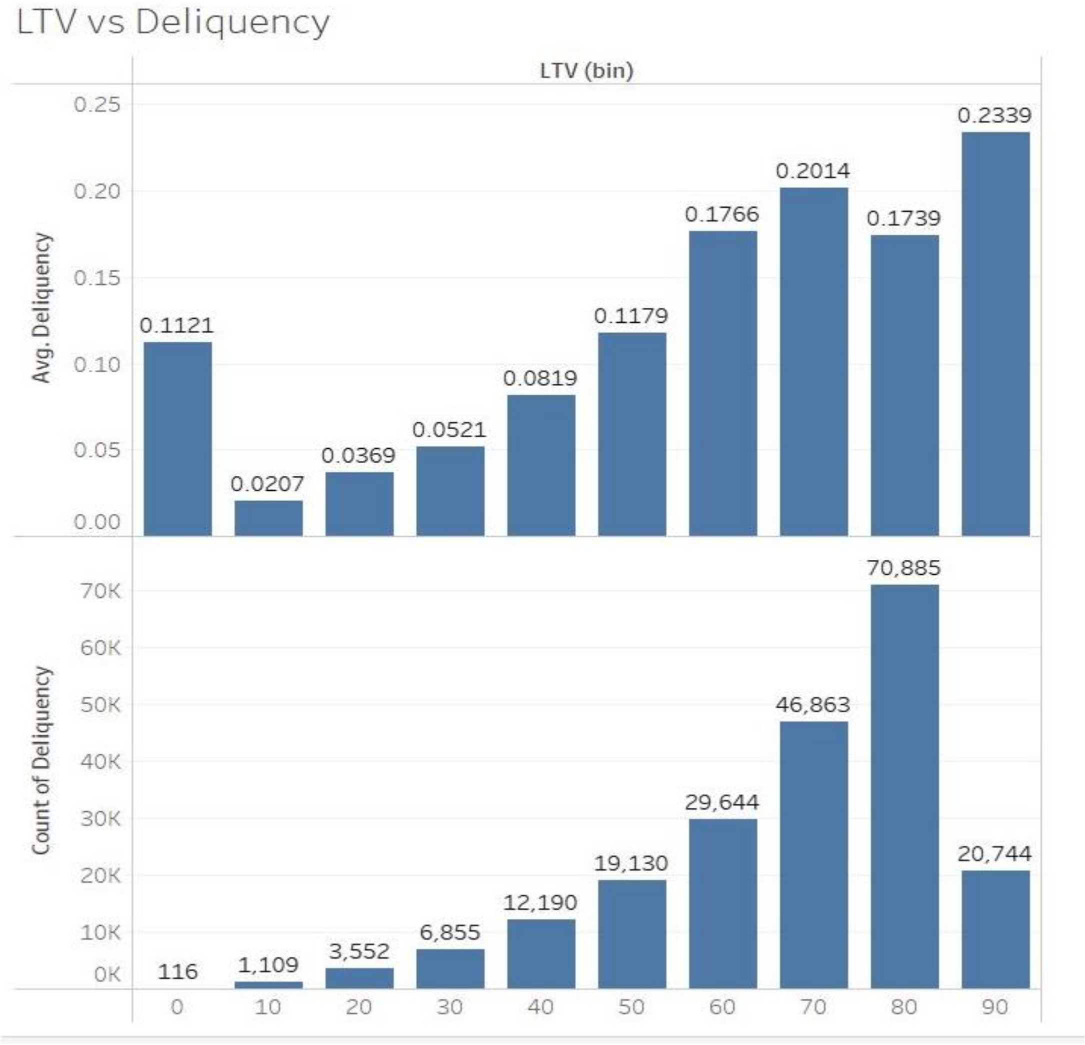
It is observed that lower the credit score more are the chances of delinquency. Customers with credit score up to 450 have 70% chances of being delinquent. As the credit score increases the chances of delinquency reduces.

2. Property State



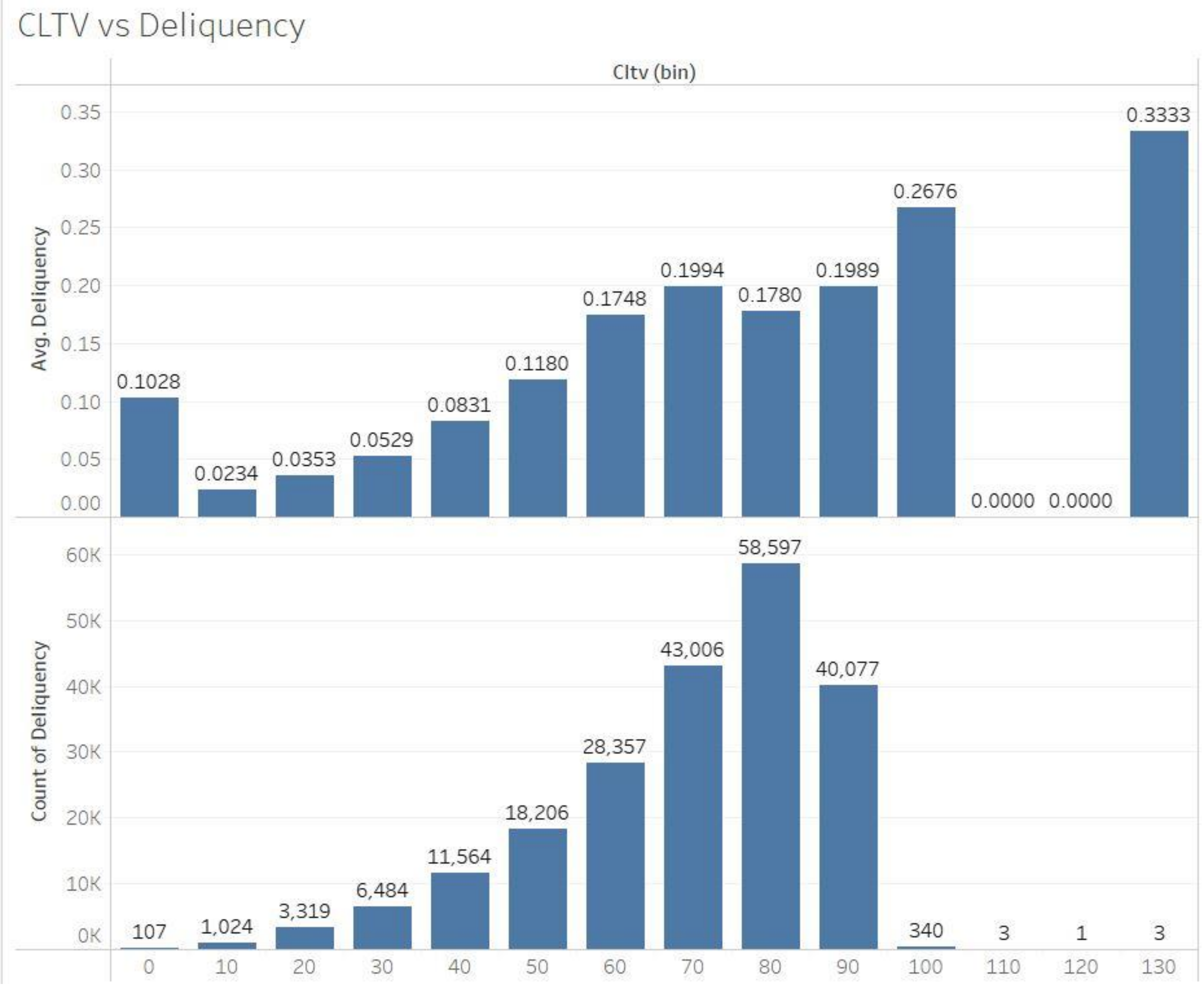
It is observed that the delinquency rate is higher in few states such as Nevada(39.59%), Florida(34.84%) and Arizona(32.17%). However, we cannot conclude based on given information any reasoning behind the same.

3. LTV(Loan-to-Value)



It is observed that as the LTV increases the delinquency rate also increases being highest at 90 for 23.39%. It is further observed that the delinquency value is less than 12% for values below 50.

4. CLTV



It is observed that as the CLTV increases the delinquency percentage also increases being highest at 100 for 26.76%. It is further observed that the delinquency value is less than 12% for values below 50.

Observed Effects	Recommendation
1. Borrower Credit Score: It is observed that lower the credit score more are the chances of delinquency. Customers with credit score up to 450 have 70% chances of being delinquent. As the credit score increases the chances of delinquency reduces.	Fannie Mae should check the credit score , as a lower credit score means that there are higher chance of delinquency. They should device a better system to grant loans to such customers so that they are able to bear such failure costs.
2. Property State It is observed that the delinquency rate is higher in few states such as Nevada(39.59%), Florida(34.84%) and Arizona(32.17%).	Nothing can be recommended based on this observation. As there are likely to be more underlying factors which needs to be evaluated to understand as to why this issue is occurring in the given regions.
3. LTV(Loan-to-Value) It is observed that as the LTV increases the delinquency rate also increases being highest at 90 for 23.39%. It is less than 12% for values below 50.	The risk is higher for customers having LTV>50 therefore Fannie Mae with the help of subject matter expert should understand and device a solution to grant loan to such customers and consider the risk involved.
4. CLTV(Combined Loan-to-Value) It is observed that as the CLTV increases the delinquency percentage also increases being highest at 100 for 26.76%. It is observed that the delinquency value is less than 12% for values below 50.	The risk is higher for customers having CLTV>50 therefore Fannie Mae with the help of subject matter expert should understand and device a solution to grant loan to such customers and consider the risk involved.

<u>Question 5:</u>

With the downfall observed in 2007, Fannie Mae should have considered the market scenario. As it was becoming easier for qualified borrowers to sell the house for any reason and lose less money by allowing bank to foreclose. As even the qualified borrowers ended up paying more than the house was worth. Mortgage-backed securities was a big cause of the financial crisis lenders who issued mortgages to customers were selling those mortgages to bigger banks for repackaging into mortgage-backed securities. In our case we can observe various Sellers that were taken in by Fannie Mae. This increased the chances of people with poor credit history to get loans easily and as observed people with lower credit history have higher delinquency. Fannie Mae should have kept a threshold value for borrower's credit history below which they should not have provided the loan.