Business Case.

Fannie Mae provides mortgages aimed at stabilizing the housing market and promoting homeownership. The goal is to find out whether a customer will be delinquent or not while paying back to Fannie Mae.

We will use a regression model, decision trees, random forest, boosted trees, and support vector machines to analyze the data. The dataset has a total of 32 features and 2,11,088 data points.

Financial implications:

Assumed Revenue			
Rate of interest	6.25 % annually	0.01	*6.25 is the avg for this dataset*
NPER	12 months * 10 years	120	
Original Principal Amount		$202,337	*Calculated from original amount field*
Start Period		1	
End Period		120	
CUMIPMT		-$70,284	
Revenue		$70,284	

Cost		
Lawyer team cost	$10,000	
Filing and Paperwork	$2,000	
Operational costs	$3,000	*People working for this project, etc*
	$15,000	

Payoff Matrix

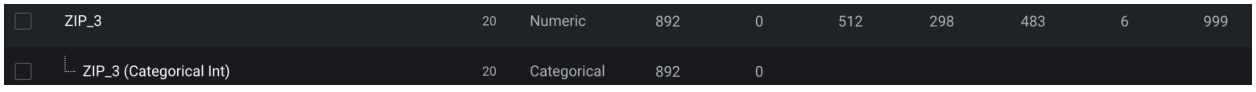
	Delinquency	Actual	Fannie Mae provides the mortgage?	Revenue	Cost	Total
TP	Yes	Yes	No	0	0	0
TN	No	No	Yes	$70,284	-$1,500 (Operational)	$68,784
FP	Yes	No	No	0	0	0
FN	No	Yes	Yes	$5,271	-$15,000	-$9,729

		Predicted	
		0	1
Actual	0	TN (68,784)	FP (0)
	1	FN (-9,729)	TP (0)

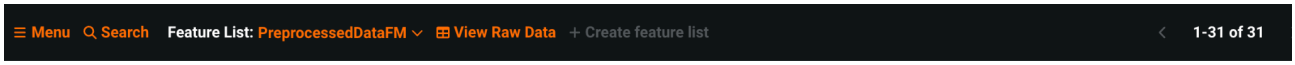
Q1. Data Preprocessing and exploratory data analysis.

- Our target variable is delinquency as we want to know whether the customer will continue to pay our interest and principal amount or not.
- Here we have excluded LoanID as it was unique for each value and didn't contribute towards the findings. Also, we changed Zipcode from numeric to categorical.
- Hence, we have taken 31 features provided in the data set for all the data points.
- We will compare logistic regression, decision trees, random forest, boosted trees and support vector machines.

Zip Code converted to categorical:

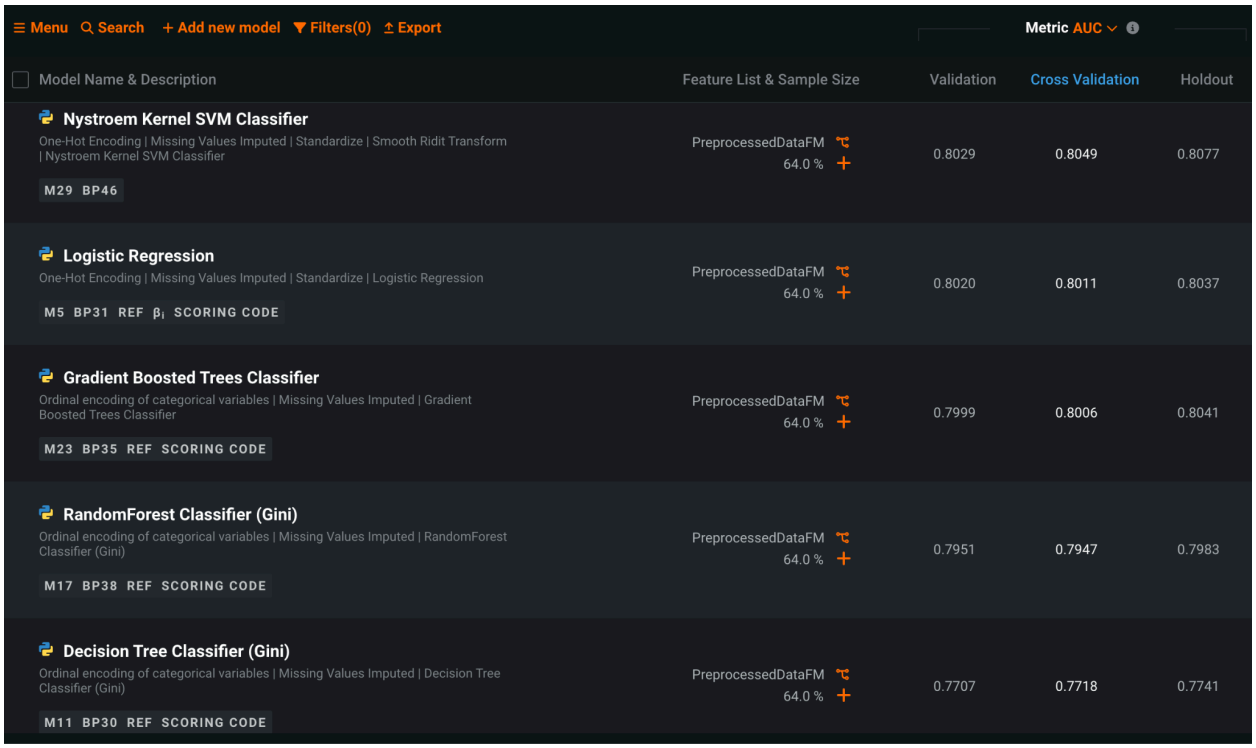


Created a new feature list without LoanID called PreprocessedDataFM.



Q2. Models and metrics.

The models we have considered here are logistic regression, decision trees, random forest, boosted trees, and support vector machines. I also know about knn but I couldn't find it on DataRobot.

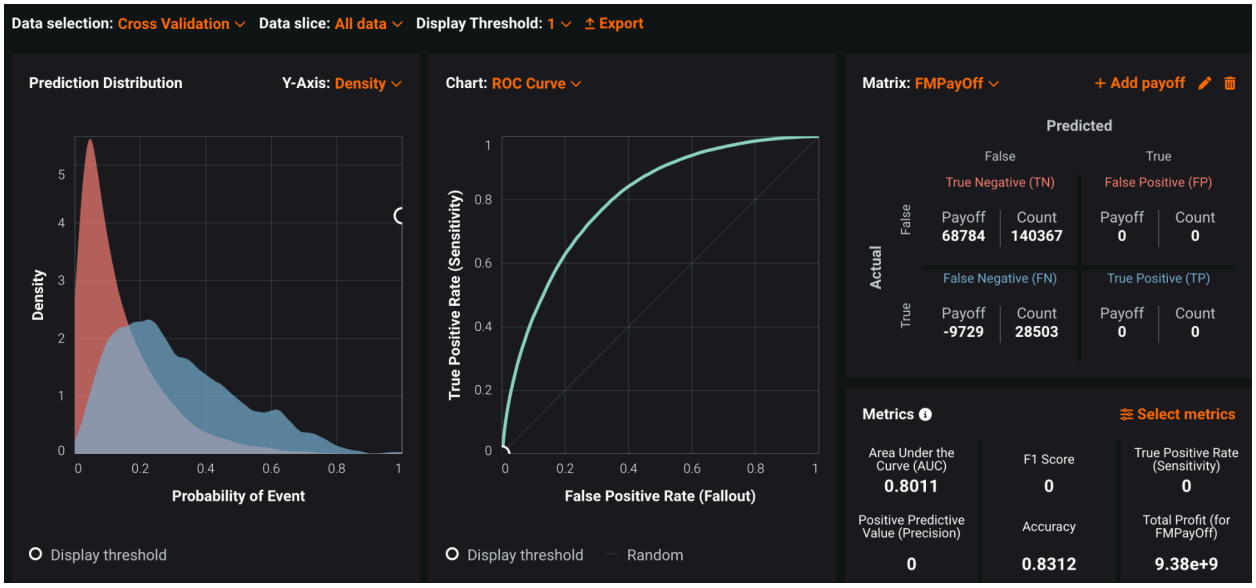


The values for AUC look good for both cross validation and holdout. It can be seen that considering AUC, SVM classifier performs the best here, and logistic regression has a very close value of AUC to it too.

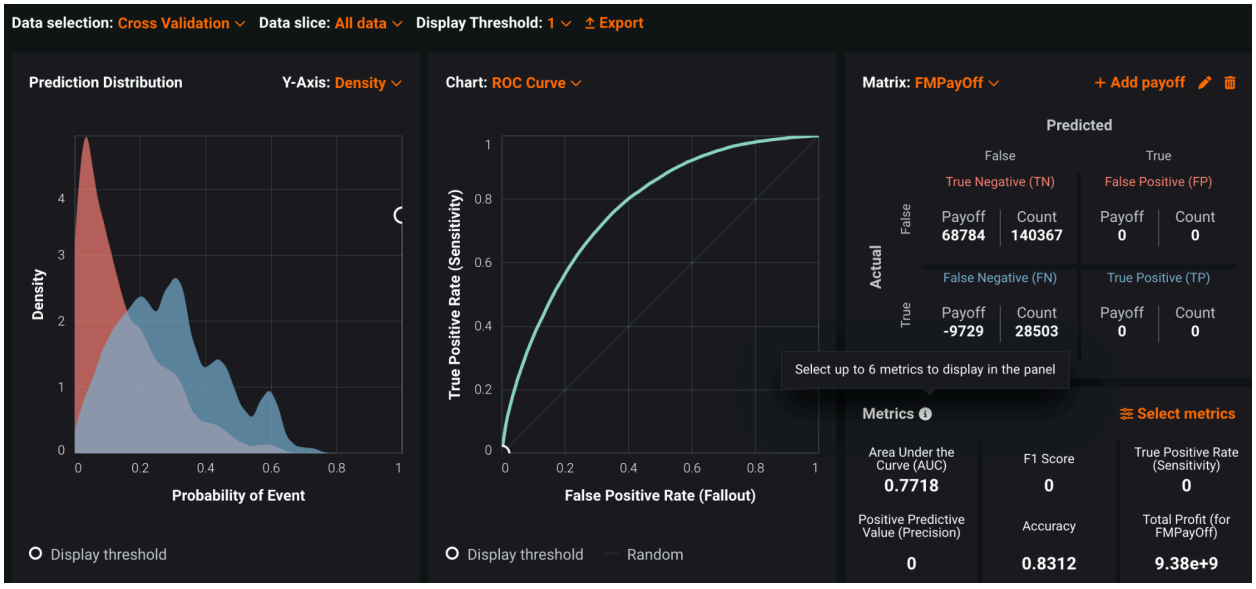
Cross Validation comparison of all models:

All the values are captured by keeping the maximum payoff.

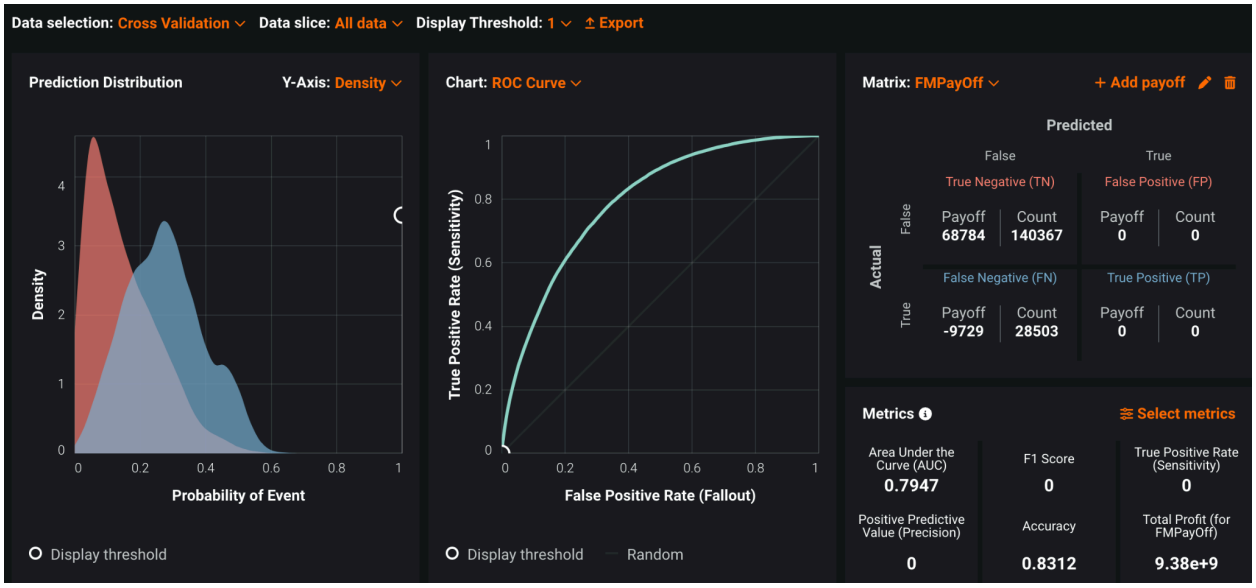
Logistic Regression



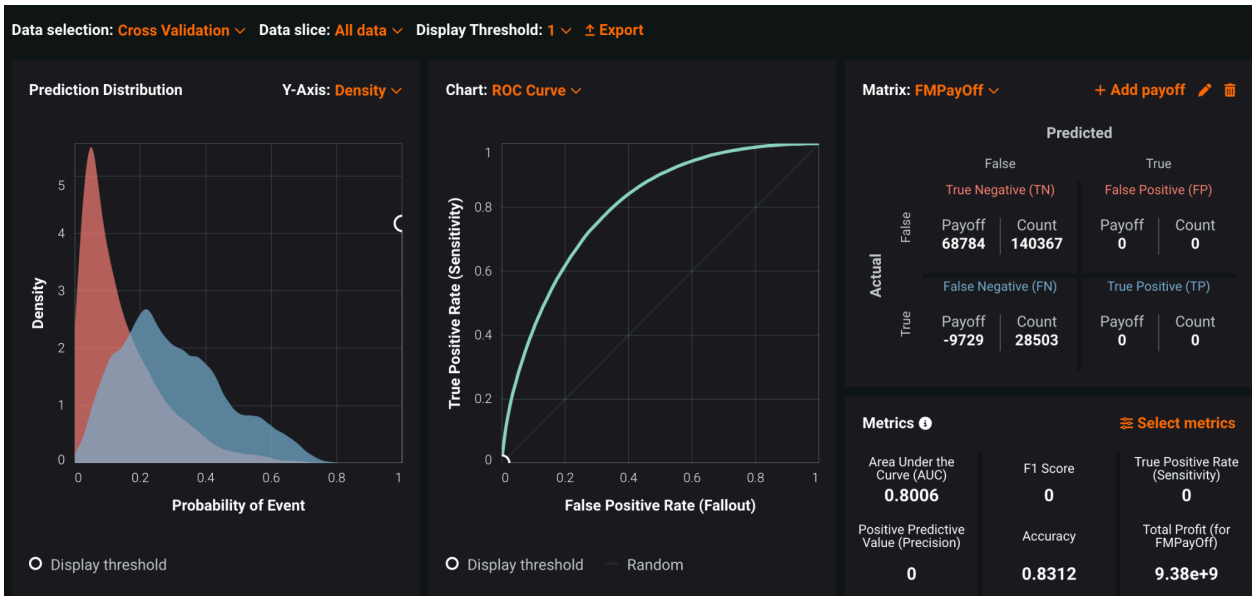
Decision Tree Classifier



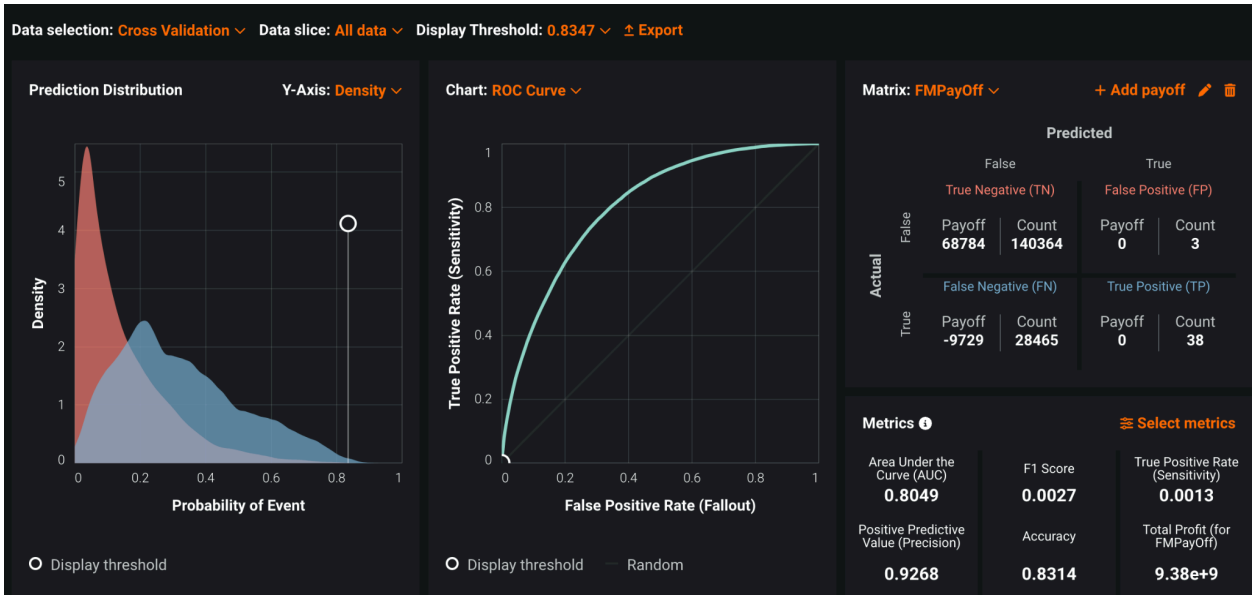
Random Forest Classifier



Boosted Trees Classifier



SVM Kernel Classifier



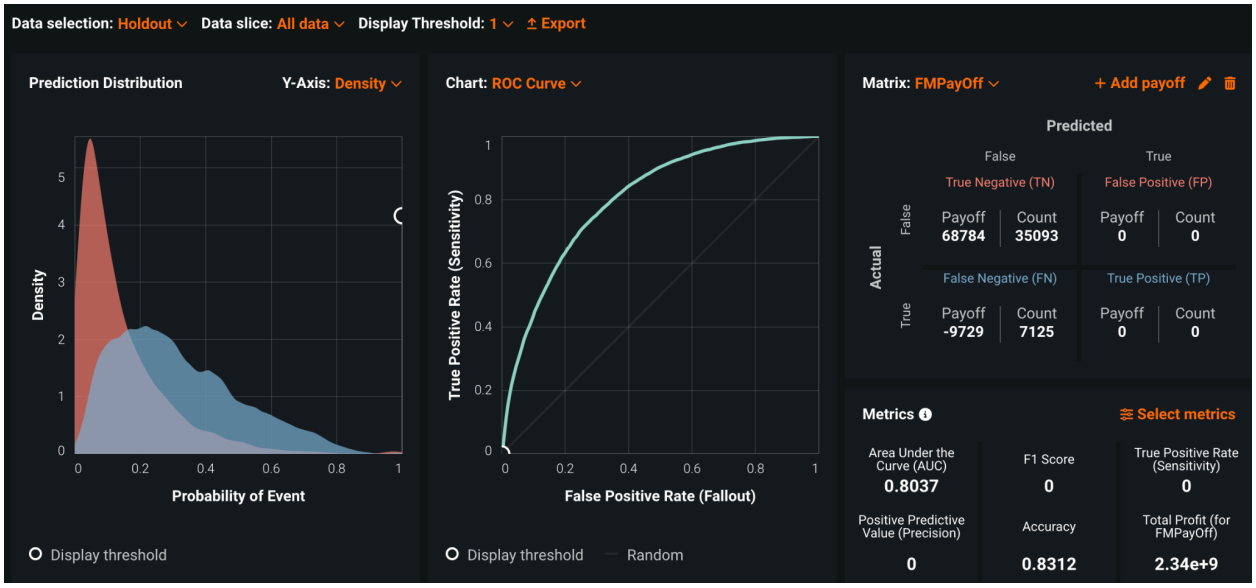
	Cross Validation Metrics at Maximum Payoff				
	Logistic Regression	*Decision Tree*	*Random Forest*	*Boosted Trees*	*SVM*
Recall	0	0	0	0	0.0013
Precision	0	0	0	0	0.9268
F1	0	0	0	0	0.0027
Accuracy	0.8312	0.8312	0.8312	0.8312	0.8314
ROC AUC	0.8011	0.7718	0.7947	0.8006	0.8049
Maximum Payoff	9.38 billion	9.38 billion	9.38 billion	9.38 billion	9.38 billion
Threshold	1	1	1	1	0.8349

For cross validation, the payoff for everyone's the same. But considering F1, SVM is doing slightly better than other models and the precision seems to be a bit high, but the recall is still very low. Overall, SVM is a better model for this case.

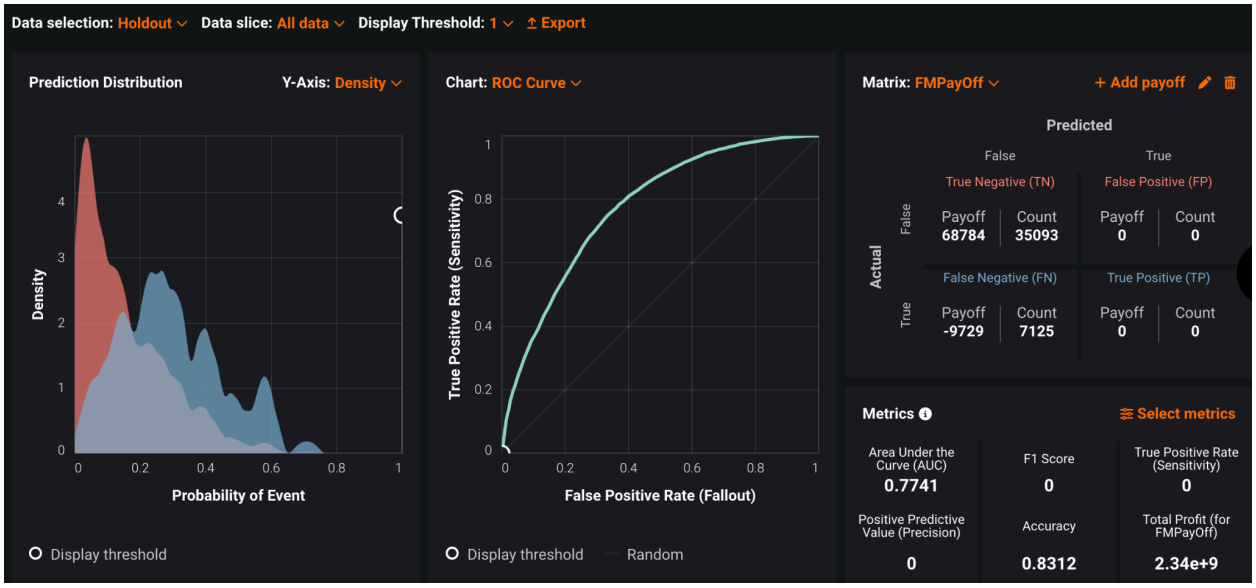
Holdout comparison of all models:

All the values are captured by keeping the maximum payoff.

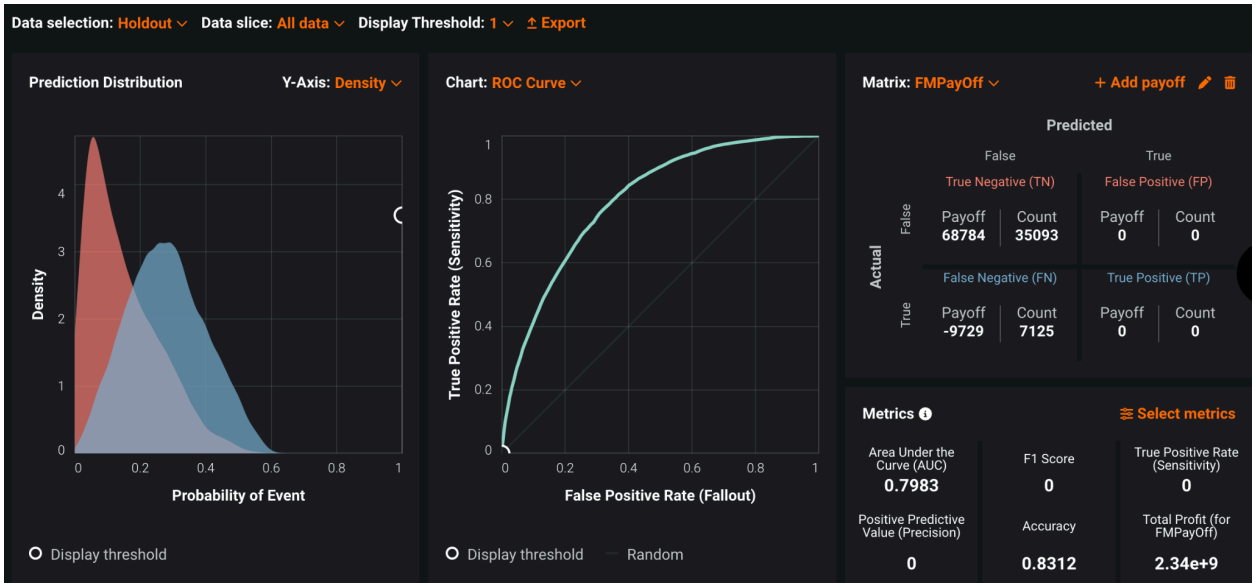
Logistic Regression



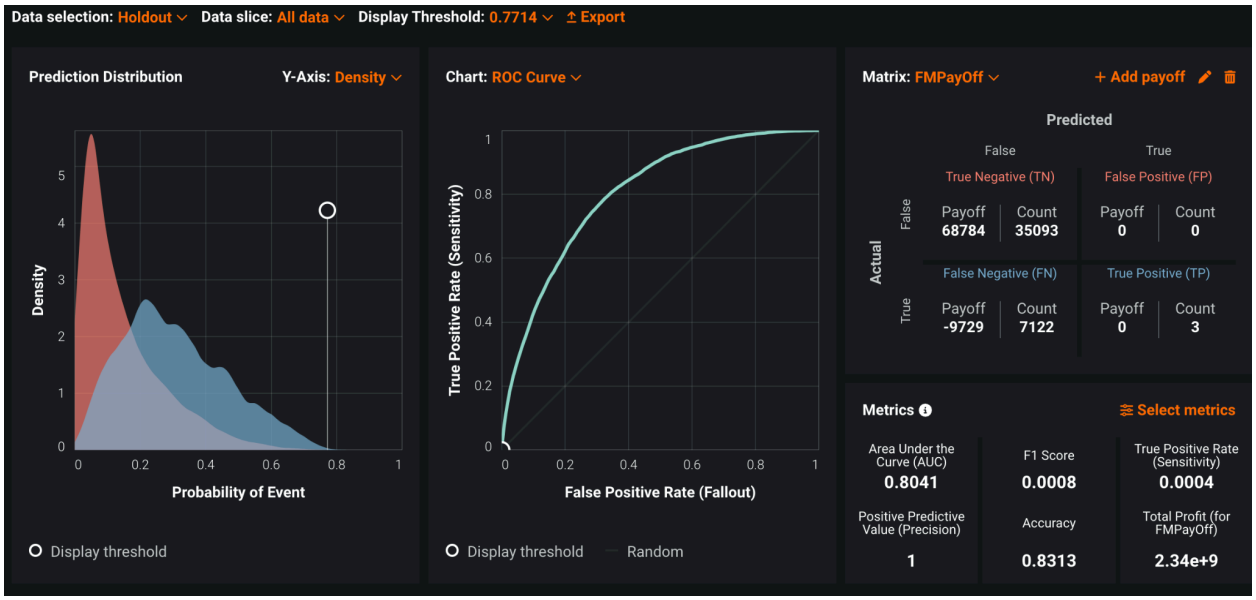
Decision Tree Classifier



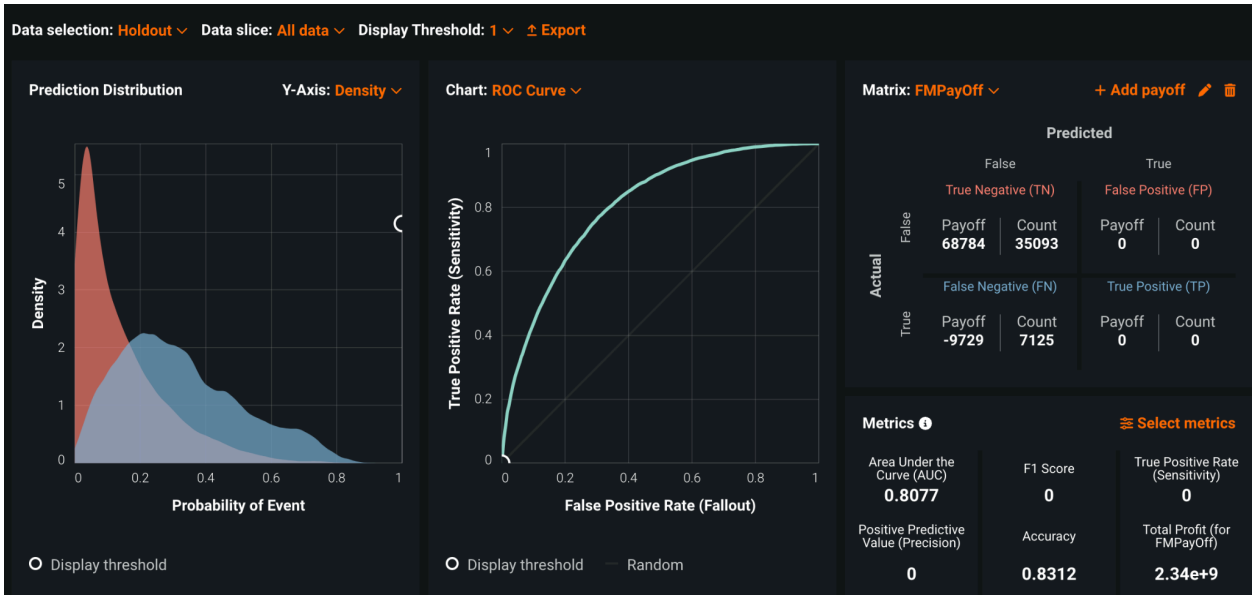
Random Forest Classifier



Boosted Trees Classifier



SVM Kernel Classifier



	Holdout Metrics at Maximum Payoff				
	Logistic Regression	*Decision Tree*	*Random Forest*	*Boosted Trees*	*SVM*
Recall	0	0	0	0.0004	0
Precision	0	0	0	1	0
F1	0	0	0	0.0008	0
Accuracy	0.8312	0.8312	0.8312	0.8313	0.8312
ROC AUC	0.8037	0.7741	0.7983	0.8041	0.8077
Maximum Payoff	2.34 billion	2.34 billion	2.34 billion	2.34 billion	2.34 billion
Threshold	1	1	1	0.7714	1

For holdout, all the models predict the same maximum pay off. Considering F1, the boosted trees model it is doing slightly better than all the other models.

Payoff matrix would be the best metric for the organization as it generates the most revenue. Also, the models have the same payoffs but Boosted trees are doing slightly better for the holdout values.

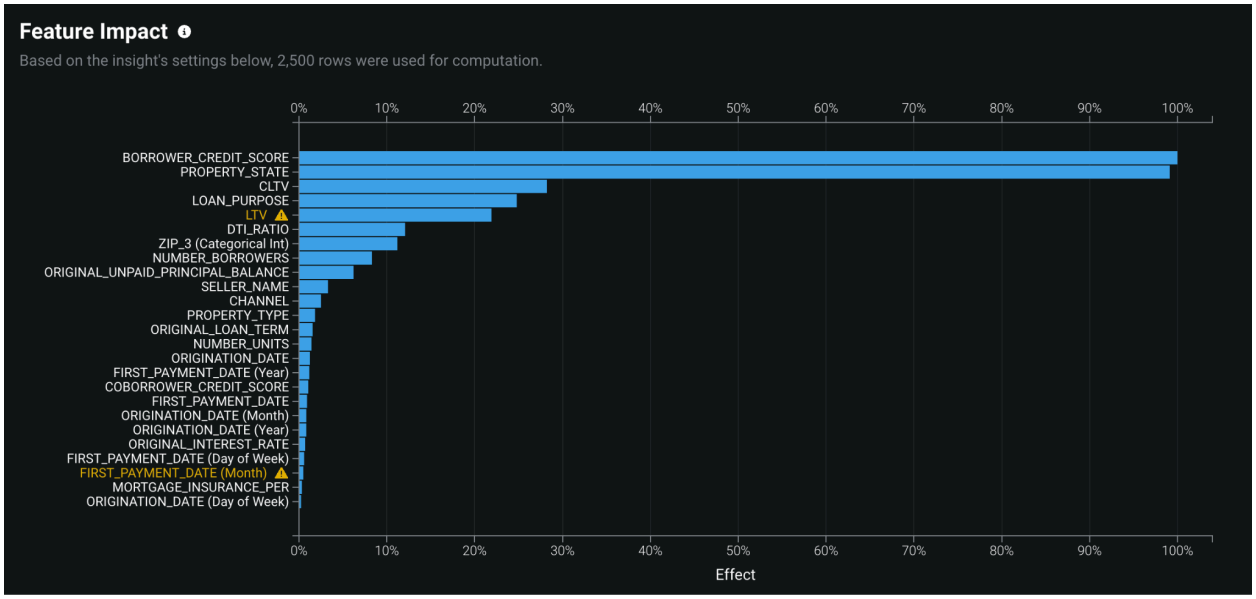
Q4. Top 4 predictors of customer churn.

The best performing model was Boosted Trees and according to the same model, the features that impact churn the most are selected.

The top 4 here are borrower_credit_score, property_state, cltv, and loan_purpose.

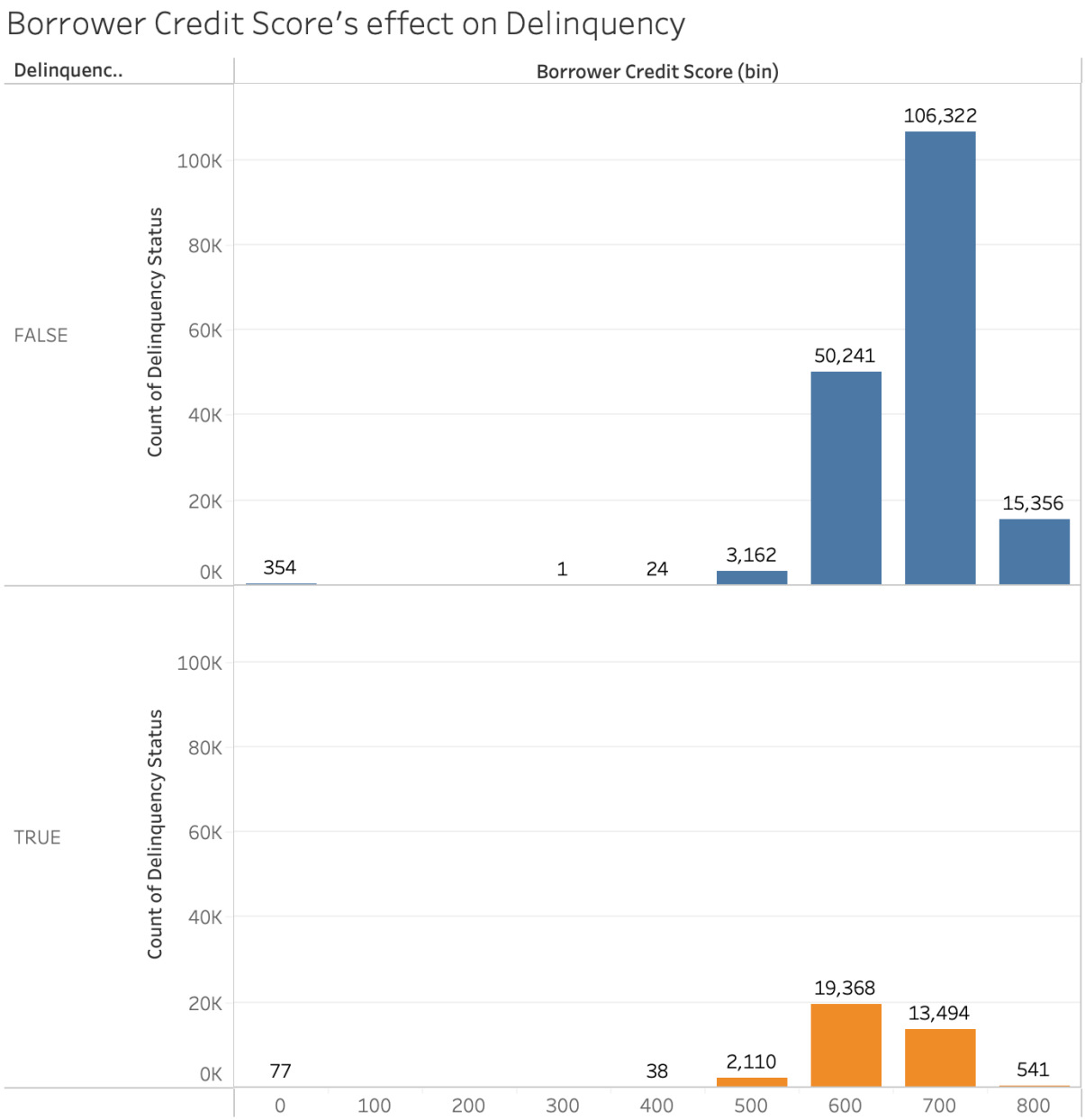
Here, we can see that there are some features which are redundant and are in high correlation with each other.

The first four features don't have any feature that is being redundant hence we consider the same four mentioned above.



Feature impact list

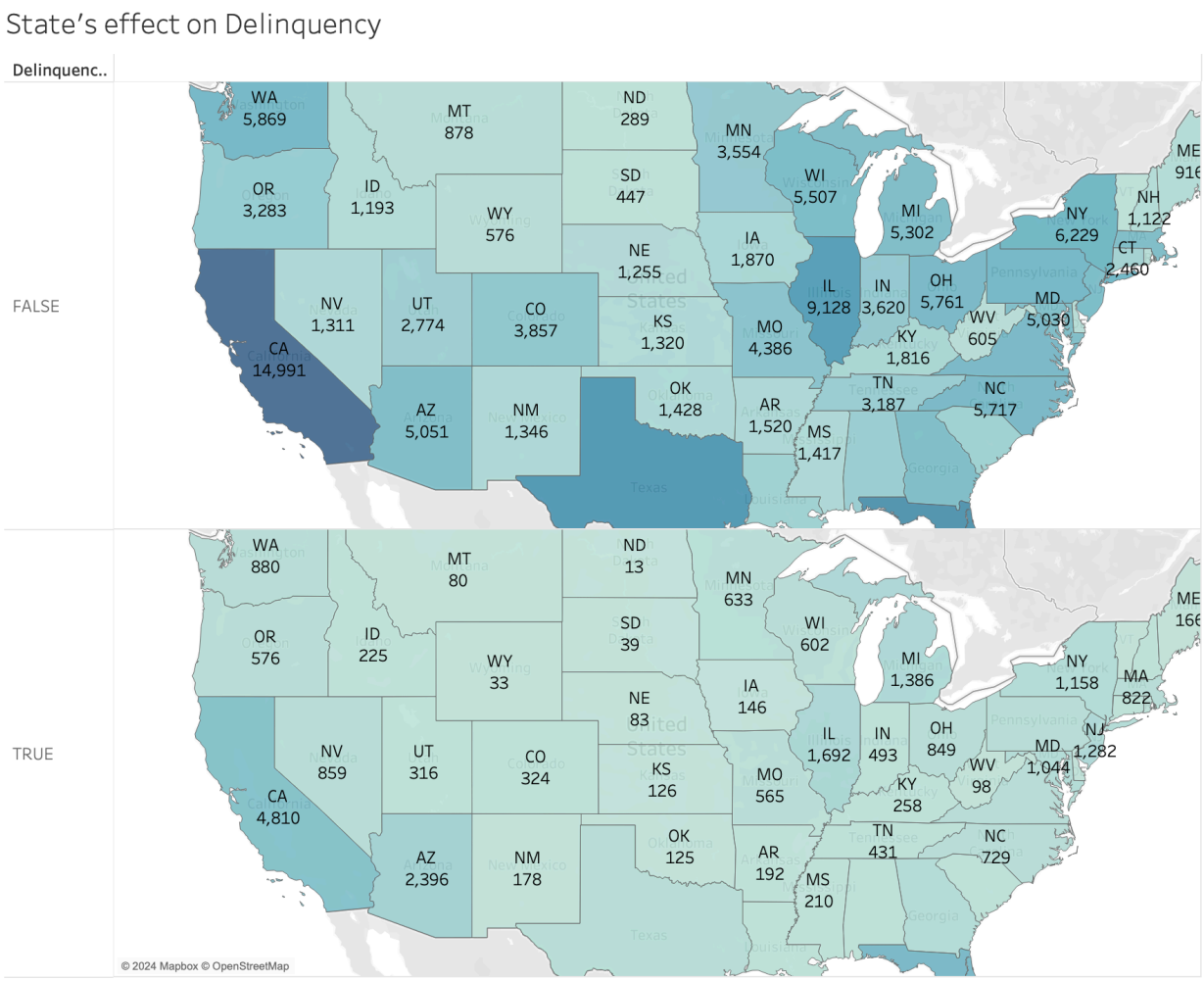
1st Feature (borrower_credit_score)



Effect : 61% of the total people whose credit score ranges from 400-500 delinquent, 40% of people whose score is from 500-600 delinquent, 27% for whose score is from 600-700 delinquent, 0.12% from the credit score of 700-800 and 3.4% from who had 800 and above. Hence, the delinquent rate is reduced for people above 700 credit score.

Recommendation: Fannie Mae may have checked the credit scores before mortgaging the properties.

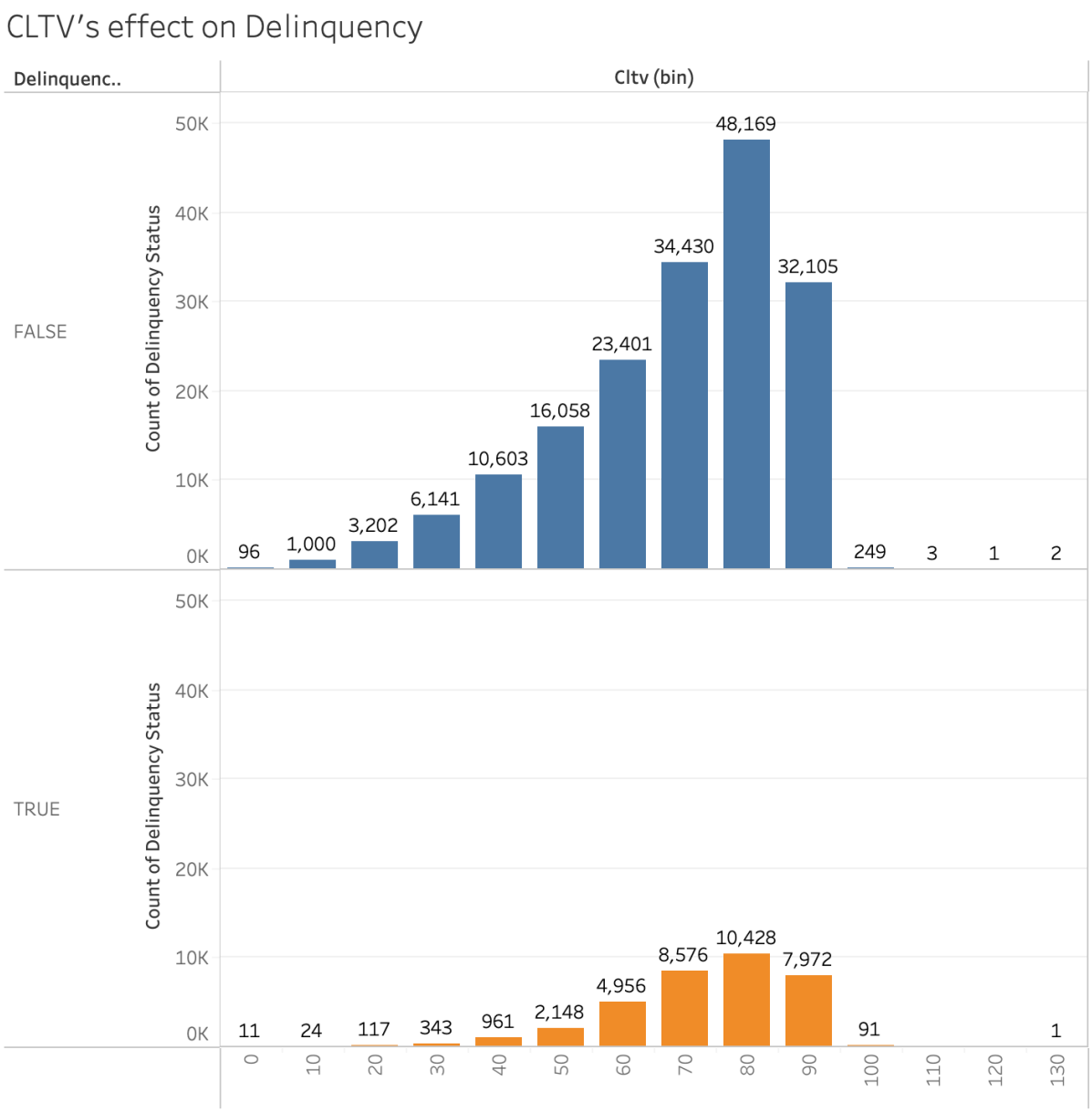
2nd Feature (property_state)



Effect : It is observed that the state of California has the most delinquency rate.

Recommendation: The states that have high delinquency rate should be checked thoroughly because the documentation may be incorrect or the evaluation wouldn't be proper though we aren't sure of the underlying reasons.

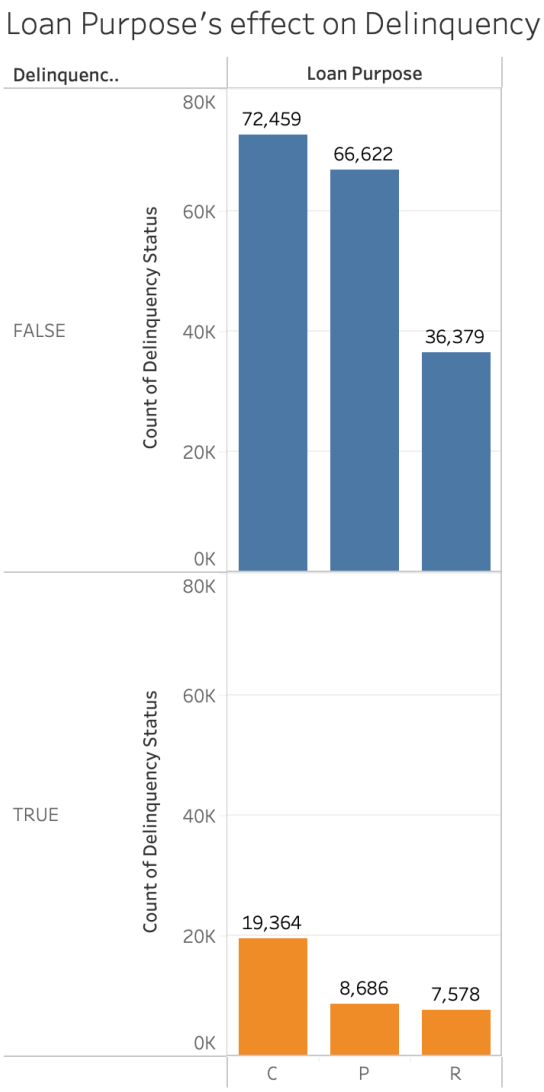
3rd Feature (CLTV - Combined Loan to Value)



Effect : The delinquent rate decreased with the increase in CLTV, and specifically after 40 which makes sense because the higher the CLTV is, the higher the chances of the loan being repaid are.

Recommendation: The CLTV score should be checked before taking the mortgage because it does make sense that the higher the CLTV is, the less chances of delinquency are there.

4th Feature (loan_purpose)



Here, P= PURCHASE, R=REFINANCE, and C=MODIFIED

Effect : 21% of modified mortgages lent by Fannie may are delinquent, 11% of purchases are delinquent, and 17% of refinanced are delinquent.

Recommendation: We aren't sure of the reasons for this, and the different types may have some reasons behind people defaulting on their mortgages.

Q5. Did Fannie Mae have information that could have accurately predicted defaults among mortgages issued in Q1 2007?

Yes, Fannie May did have some factors that it could have considered and made better decisions of providing the mortgages on Q1 of 2007 like CLTV, and Borrower's credit score. Taking a look at the other factors would have also provided insights and helped Fannie Mae.