

## Case Study – 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. The objective was to identify people who might default on the loan, if sanctioned, by Fannie Mae for a particular property. A model was required to predict the best suitable borrowers for the loan who would repay the loan on time, thereby making profit for the company through the interest charged by them. Fannie Mae would want to make more informed decisions as to who to sanction the loan to minimize the number of defaulters thereby in turn maximizing the profit. The dataset used for modeling is from 2007 (quarter 1) consisting of 211088 entries.

### Payoff Matrix –

Average house price = \$188,000

Average loan Amount = 80% of average house price (assuming the down payment made to be 20% of the house price) =  $0.80 * 188000 = \$150,400$

Average loan repayment tenure = 10 years =  $12 * 10 = 120$  months

Average rate of interest = 6%

Average interest earned by Fannie Mae = 0.5%

Assumptions for defaulting customers –

1. The property is sold after 2 years.
2. The property is sold for 85% of the original cost = \$159800
3. The property was refurbished by Fannie Mae after eviction for an approximate cost of \$25,000.
4. Fannie Mae incurs an additional legal cost of approximately \$10,000 for filing paperwork against the defaulter.

True Positive – Model predicts that the customer is likely to default and hence Fannie Mae doesn't acquire the property. There is no cost and no revenue as the loan is not issued.

TP = 0

True negative – The model predicts that the customer would repay and hence Fannie Mae provides the loan. The company makes revenue as they acquire the property.

TN = \$3822.6875 = approx. \$3823

<b>Rate</b>	<b>0.00041667</b>
<b>Nper</b>	<b>120</b>
<b>Pv</b>	<b>150400</b>
<b>Start_period</b>	<b>1</b>
<b>End_period</b>	<b>120</b>
<b>Type</b>	<b>0</b>
<b>CUMIPMT</b>	<b>-3822.6875</b>

False Positive – Model predicts a potential customer to be a defaulter. Fannie Mae doesn't sanction the loan and hence no revenue is generated. There is no associated cost also. Although it might be considered as missed opportunity cost.

FP = 0

False Negative – Model incorrectly predicts the customer as non-defaulter and hence Fannie Mae makes the loan offer. The customer is a defaulter and hence Fannie Mae now must bear the loss. Assumption – Customer does no repayment and defaults from the 1<sup>st</sup> month itself.  
 FN = \$-35,180

Loss = Price at which the property is sold – Loan Amount – interest rate for 2 years – cost for refurbishing the property – costs incurred from the legal actions  
 Loss = 159800 - 150,400 – 9579.6748 - 25,000 – 10,000 = - 35179.6748 = approx. -35180

<b>Rate</b>	<b>0.005</b>
<b>NPER</b>	<b>24</b>
<b>PV</b>	<b>150400</b>
<b>Start_period</b>	<b>1</b>
<b>End_period</b>	<b>24</b>
<b>Type</b>	<b>0</b>
<b>CUMIPMT</b>	<b>-9579.6748</b>

Final Payoff matrix –

	<b>Predicted 0</b>	<b>Predicted 1</b>
<b>Actual 0</b>	3823	0
<b>Actual 1</b>	-35180	0



# 1. Explore and pre-process data as needed. Provide a bullet list of pre-processing steps.

Pre-Processing –

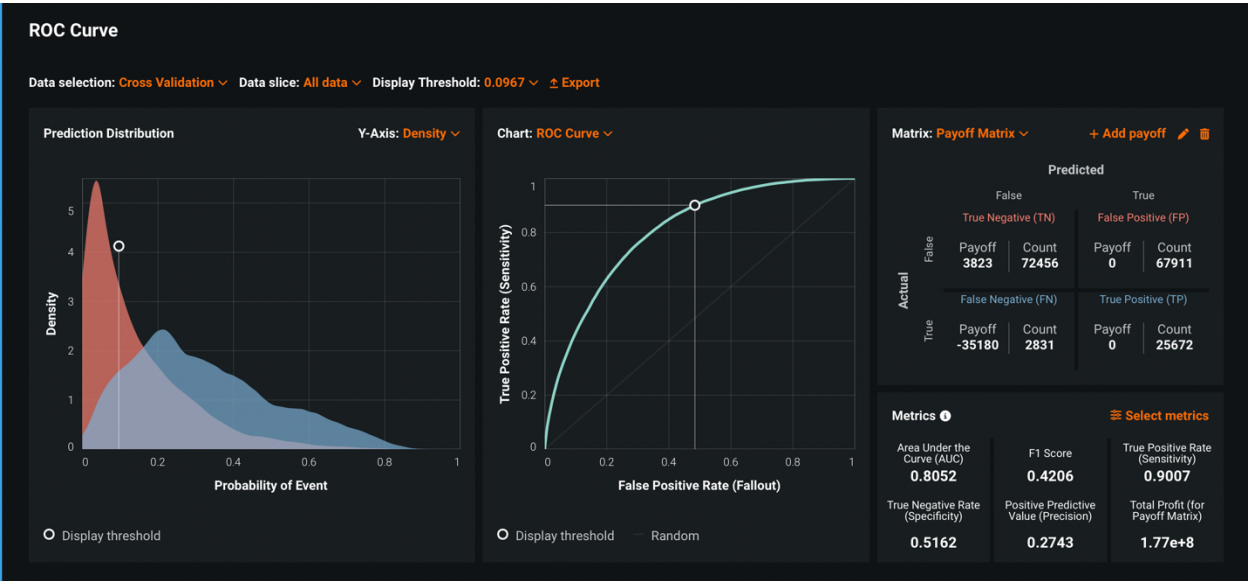
- Loan\_id field has been removed as id isn't an informative feature for modeling and it has high cardinality.
- FIRST\_PAYMENT\_DATE feature has been automatically converted by Datarobot into FIRST\_PAYMENT\_DATE (Day of Week), FIRST\_PAYMENT\_DATE (Month), FIRST\_PAYMENT\_DATE (Year), and FIRST\_PAYMENT\_DATE (Day of Month) by extracting all the essential information and hence the FIRST\_PAYMENT\_DATE feature is excluded, and these newly created features are used.
- ORIGINATION\_DATE feature has been automatically converted by Datarobot into ORIGINATION\_DATE (Day of Week), ORIGINATION\_DATE (Month), ORIGINATION\_DATE (Year), and ORIGINATION\_DATE (Day of Month) by extracting all the essential information and hence the ORIGINATION\_DATE feature is excluded, and these newly created features are used.
- Among the newly created features above we have Day of month (ORIGINATION\_DATE (Day of Month) and FIRST\_PAYMENT\_DATE (Day of Month)) which has only 1 unique value, so we remove this.
- Product\_type field has only one unique value for all the records and hence it provides no additional information for modeling, so we remove this field.
- Zip\_3 field for zip code has been converted to categorical feature.

2. Make use of all the modeling techniques that you know to build models to forecast mortgage delinquency.

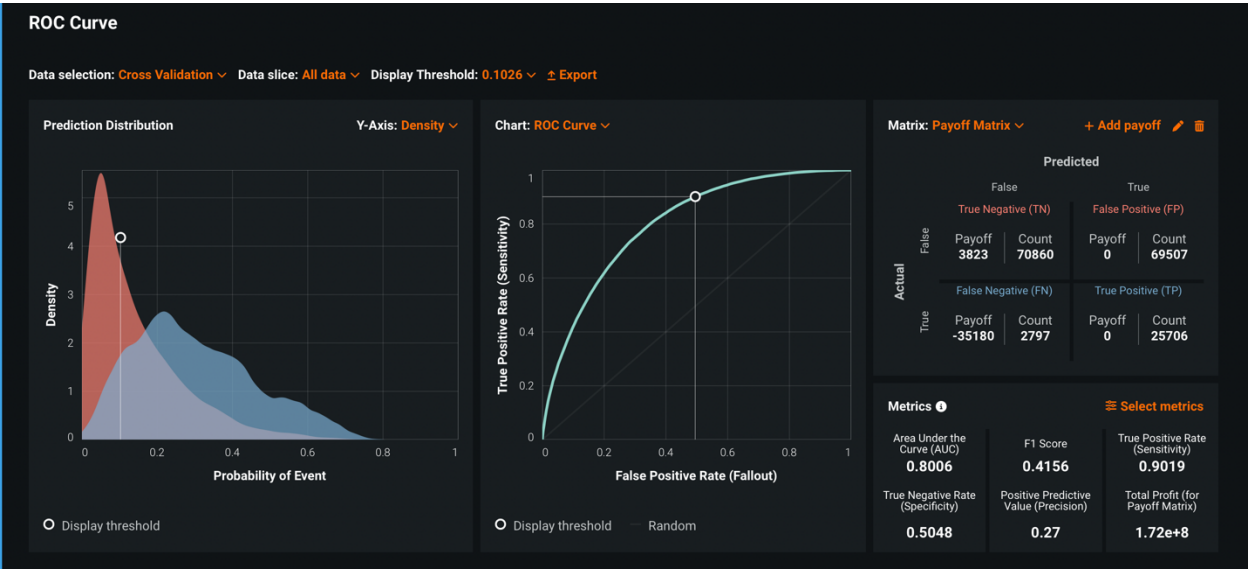
The below models have been trained on the data –

<div><div> <b>Nystroem Kernel SVM Classifier</b></div><div>One-Hot Encoding   Missing Values Imputed   Standardize   Smooth Redit Transform   Nystroem Kernel SVM Classifier</div></div> <div>preprocessed_features 64.0 % + 0.3657 0.3642</div> <div>M24 BP46</div>	
<div><div> <b>Gradient Boosted Trees Classifier</b></div><div>Ordinal encoding of categorical variables   Missing Values Imputed   Gradient Boosted Trees Classifier</div></div> <div>preprocessed_features 64.0 % + 0.3695 0.3689</div> <div>M12 BP35 REF SCORING CODE</div>	
<div><div> <b>Logistic Regression</b></div><div>One-Hot Encoding   Missing Values Imputed   Standardize   Logistic Regression</div></div> <div>preprocessed_features 64.0 % + 0.3728 0.3719</div> <div>M18 BP31 REF <math>\beta_i</math> SCORING CODE</div>	
<div><div> <b>RandomForest Classifier (Gini)</b></div><div>Ordinal encoding of categorical variables   Missing Values Imputed   RandomForest Classifier (Gini)</div></div> <div>preprocessed_features 64.0 % + 0.3744 0.3743</div> <div>M30 BP38 REF SCORING CODE</div>	
<div><div> <b>Decision Tree Classifier (Gini)</b></div><div>Ordinal encoding of categorical variables   Missing Values Imputed   Decision Tree Classifier (Gini)</div></div> <div>preprocessed_features 64.0 % + 0.3869 0.3895</div> <div>M6 BP30 REF SCORING CODE</div>	

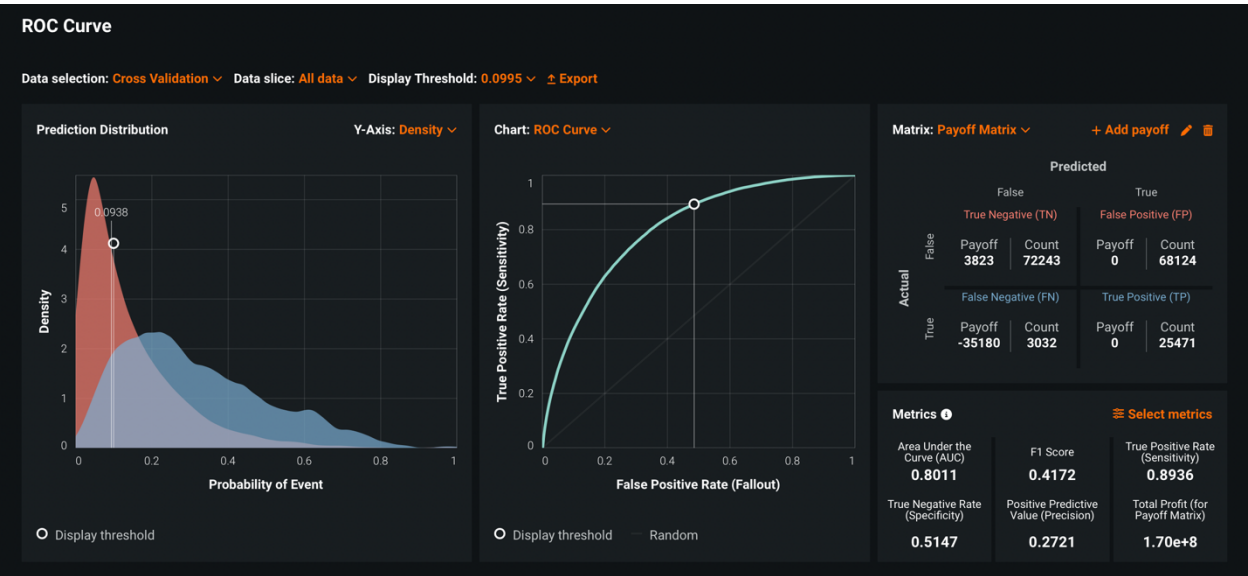
a. Nystroem Kernel SVM Classifier



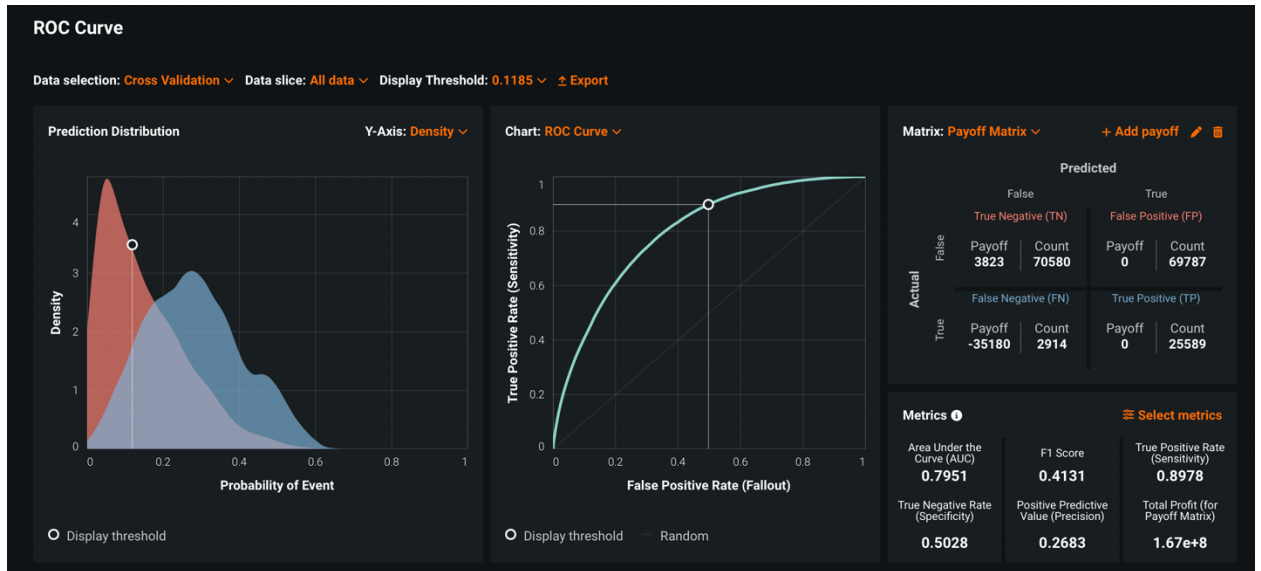
b. Gradient Boosted Trees



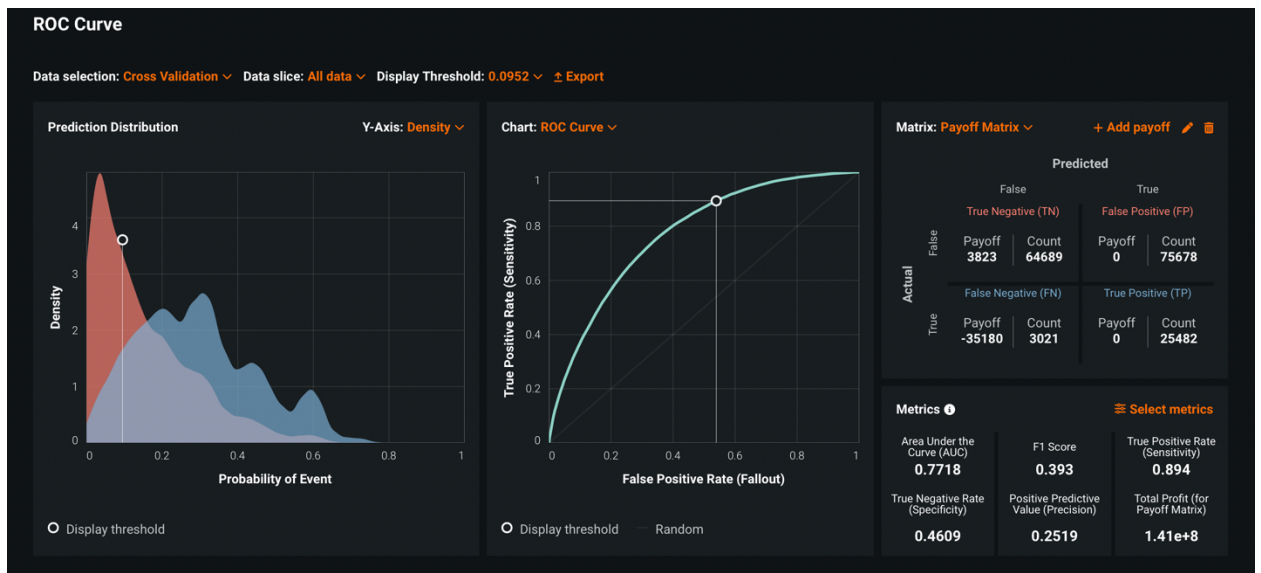
c. Logistic Regression



d. RandomForest Classifier (Gini)



e. Decision Tree Classifier (Gini)



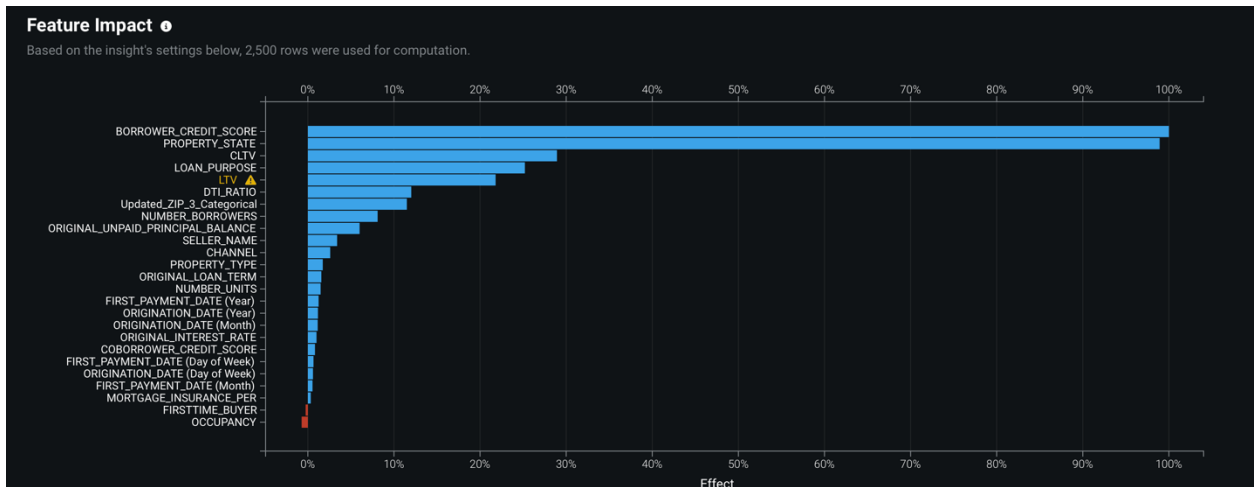
**3. Show model performance metrics: Recall, Precision, Specificity, F1, ROC AUC, max payoff. What is the best metric to evaluate model performance?**

All the below values for the models are noted at the threshold that maximizes the profit -

<b>Models</b>	<b>Recall</b>	<b>Precision</b>	<b>Specificity</b>	<b>F1</b>	<b>ROC AUC</b>	<b>Max payoff</b>
<b>Nystroem Kernel SVM Classifier</b>	0.9007	0.2743	0.5162	0.4206	0.8052	177,000,000
<b>Gradient Boosted Tree Classifier</b>	0.9019	0.27	0.5048	0.4156	0.8006	172,000,000
<b>Logistic Regression</b>	0.8936	0.2721	0.5147	0.4172	0.8011	170,000,000
<b>RandomForest Classifier (Gini)</b>	0.8978	0.2683	0.5028	0.4131	0.7951	167,000,000
<b>Decision Tree Classifier (Gini)</b>	0.894	0.2519	0.4609	0.393	0.7718	141,000,000

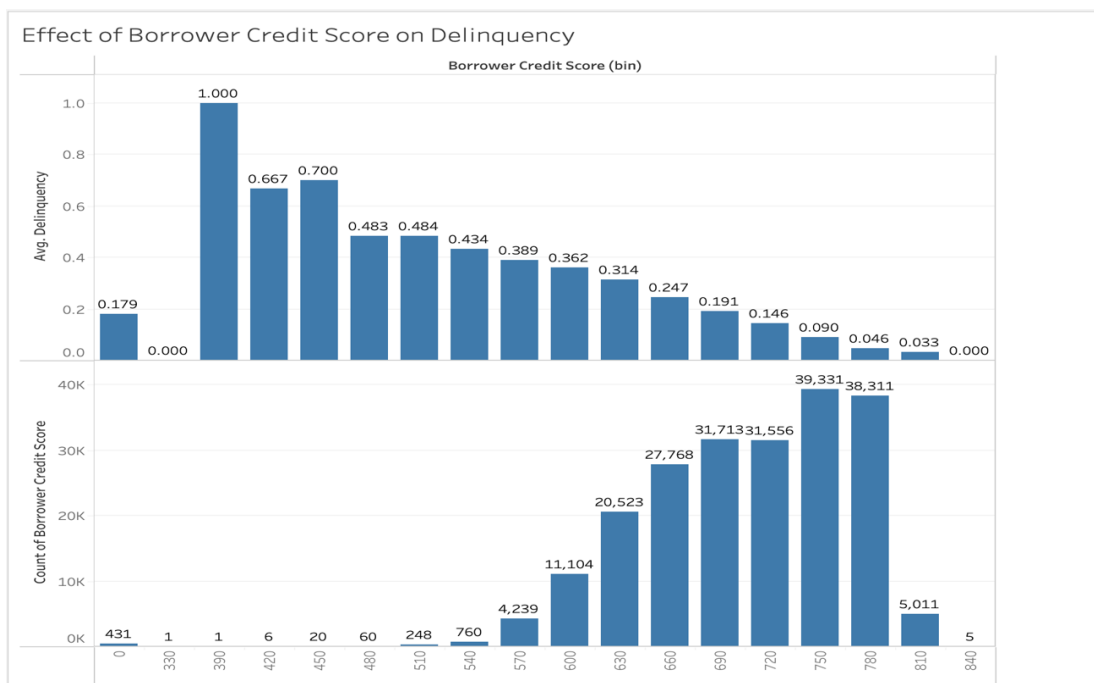
The best metric to evaluate the model performance is the maximum payoff metric. Maximum payoff metric assigns costs to the different scenarios (true positive/true negative/false positive/ false negative) thereby helping in understanding the overall profit or loss incurred by the company for each case. As the business case requires us to identify potential customers who wouldn't default on loan repayment and hence help Fannie Mae make profit, payoff metric is the best for evaluating the different models. According to the payoff metric, the Nystroem Kernel SVM Classifier is the best performing model for our business case as it generates the maximum payoff of \$177,000,000.

4. Visualize the effects of the top 4 predictors of mortgage defaults. Summarize the observed effects in 1-2 sentences for each predictor. Provide an actionable recommendation based on each observation. If the observation is not actionable, state so.



The top 4 predictors are – Borrower\_Credit\_score, Property\_state, CLTV (Combined Loan to Value) and Loan\_Purpose

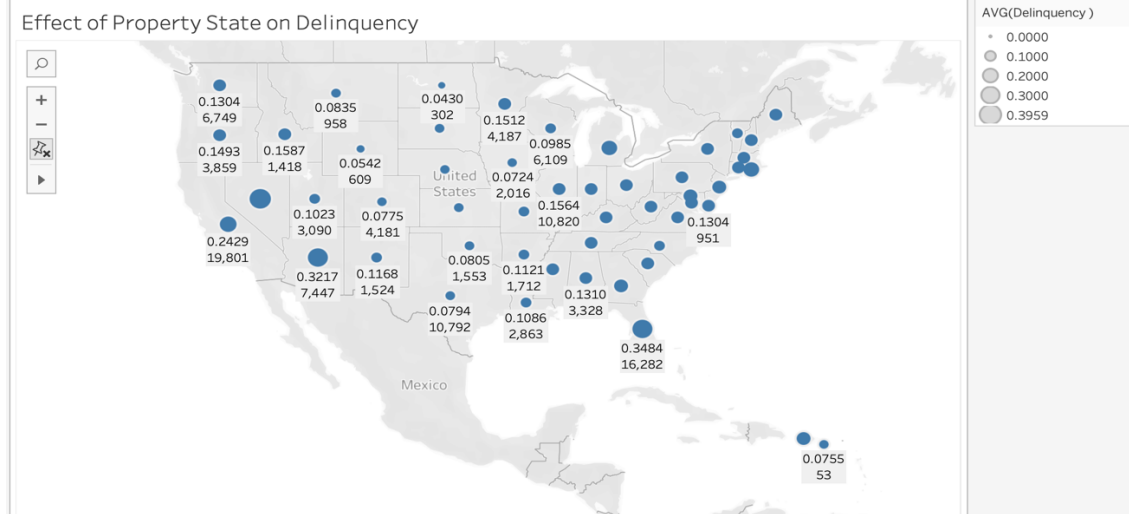
- a. Effect of Borrower Credit Score predictor



It is observed that people with lower credit scores are more likely to default on the loans and the percentage decreases as the credit score of the customer increases. People with credit scores up to 450 almost have 70% chances of being delinquent. However, for people with credit scores as high as 780, the rate of delinquency is just about 4.6% even though the number of records for such customers is much larger.

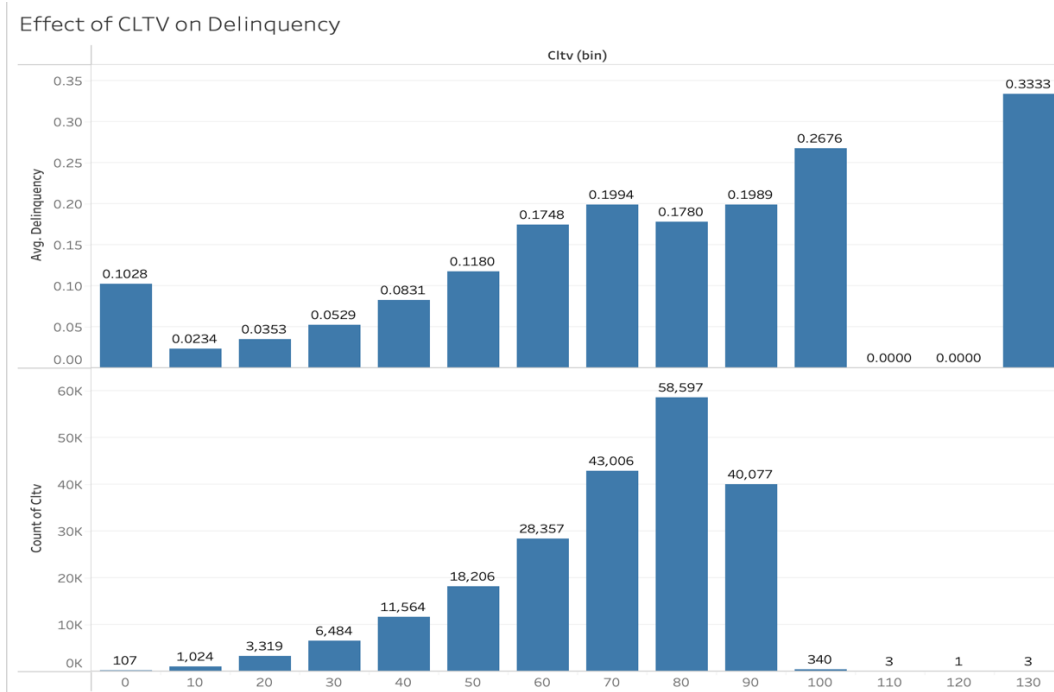


## b. Effect of Property State predictor



Certain property states have higher delinquency rates as compared to others. Though we cannot see or understand any specific pattern as such but some states like Florida (34.84%), Arizona (32.17%) and Nevada (39.59%) have higher rate of defaulters as compared to others.

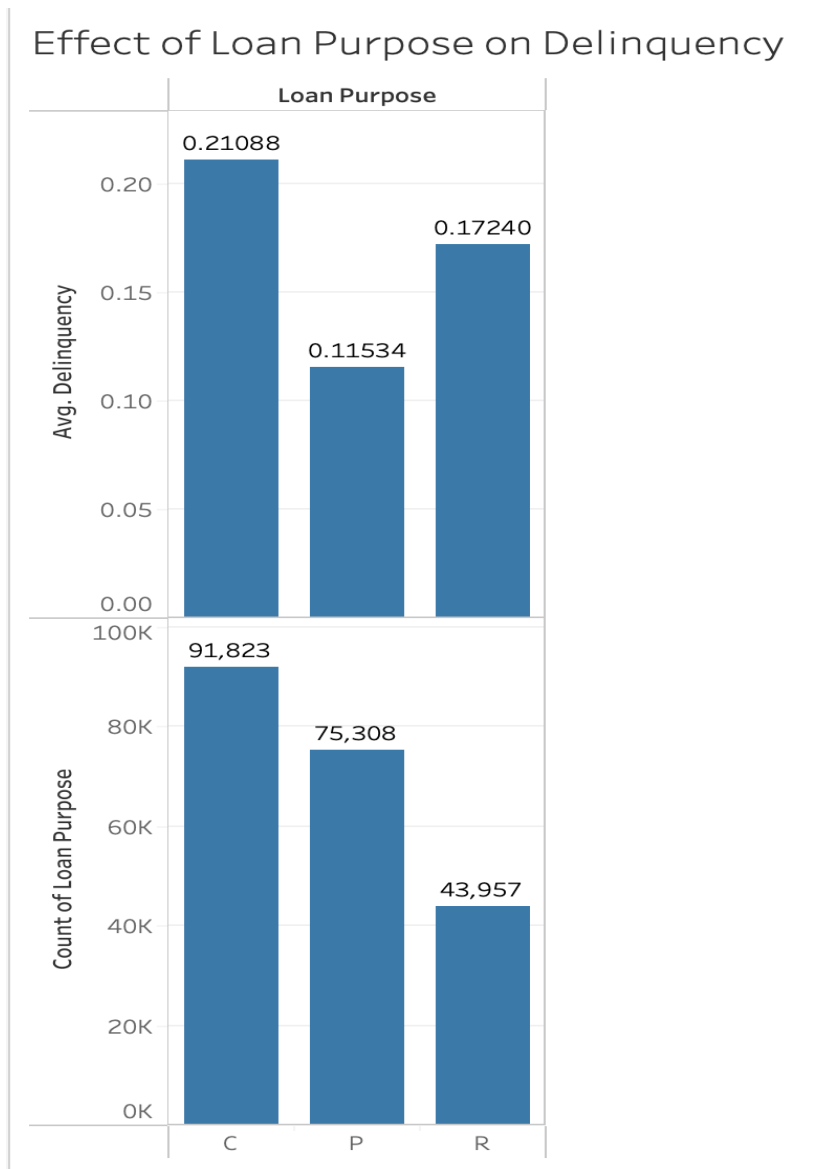
## c. Effect of Combined Loan to Value (CLTV) predictor



It is observed that the delinquency rate increases with the increase in the value of CLTV. For values below 50 the delinquency rate is not very high (almost less than 8.5%) but it goes on to increase to 26.76% for CLTV values up to 100. The maximum delinquency rate is 33.33% for CLTV value of 130 but we only have 3 records for that.



d. Effect of Loan Purpose predictor



(The data dictionary presents values for loan purpose as purchase, refinance and modified. Below observation assumes that P is Purchased, R is refinanced, and C is modified)

The delinquency rate seems to be the highest (21%) for the modified loan purpose type. Also, even though the number of customers with loan purpose type as refinanced is comparatively lower (43,957) it still has a higher default rate of 17.2% as compared to loan purpose type of purchased (11.5%)

Feature	Observed Effect	Recommendation
Borrower Credit Score	It is observed that people with lower credit scores are more likely to default on the loans and the percentage decreases as the credit score of the customer increases. People with credit scores up to 450 almost have 70% chances of being delinquent. However, for people with credit scores as high as 780, the rate of delinquency is just about 4.6% even though the number of records for such customers is much larger.	Fannie Mae should make decisions only after verifying the credit score of the customer as people with lower credit score are more likely to default.
Property State	Certain property states have higher delinquency rates as compared to others. Though we cannot see or understand any specific pattern as such but some states like Florida (34.84%), Arizona (32.17%) and Nevada (39.59%) have higher rate of defaulters as compared to others.	There can be no specific actionable recommendation based on the observation for this. Fannie Mae maybe needs to investigate this and check for possible reasons.
CLTV	It is observed that the delinquency rate increases with the increase in the value of CLTV. For values below 50 the delinquency rate is not very high (almost less than 8.5%) but it goes on to increase to 26.76% for CLTV values up to 100. The maximum delinquency rate is 33.33% for CLTV value of 130 but we only have 3 records for that.	As the rate of defaulters is higher for customers with CLTV values greater than 50, Fannie Mae should explore this segment of customers to understand the actual problem to devise a solution for the same.
Loan Purpose	The delinquency rate seems to be the highest (21%) for the modified loan purpose type. Also, even though the number of customers with loan purpose type as refinanced is comparatively lower (43,957) it still has a higher default rate of 17.2% as compared to loan purpose type of purchase (11.5%)	Fannie Mae can concentrate more on the segment of the customers with loan purpose type as modified and refinanced to reduce the default rate in these segments. They can have domain specialist to understand the reasoning for the same and then take proper action based on the insights.

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

No, Fannie Mae didn't have the information that could have helped it to accurately predict defaults among the mortgages. The accuracy for the best model is 58% which means it was able to correctly identify only 58% of the defaulters. Making predictions for defaulters in a financial institution is not an easy task and hence this information alone was not enough for the company to correctly predict defaulters.