Business Case

Fannie Mae, a key player in the US housing finance sector, has historically grappled with the complexities of risk management, particularly the prediction and mitigation of mortgage delinquencies—a matter accentuated by the 2007 housing market crash. In this predictive analysis, advanced modeling techniques like Logistic Regression, Gradient Boosted Trees, Random Forest, and Nystroem Kernel SVM will be utilized. Each model is tailored to parse through data, discerning the intricate factors that forewarn of potential delinquencies. The preparatory stage of this project entails an analysis of the data to pinpoint patterns linked to delinquencies, followed by extensive preprocessing to prime the data for modeling. The models' performances will be evaluated based on their accuracy, as denoted by metrics like Recall, Precision, F1 Score, and ROC AUC. However, a financial lens is also crucial, so the Maximum Payoff metric will quantify the economic impact of each model's delinquency predictions. This business case goes beyond just data analysis; it's a strategic initiative aimed at enhancing Fannie Mae's approach to mortgage lending. By accurately predicting which loans may default, Fannie Mae aims not only to secure its own financial standing but also to contribute to the overall stability of the housing market, reinforcing consumer confidence and advocating for responsible homeownership. The result of implementing these predictive models is a more fortified, insightful approach to mortgage lending, equipping Fannie Mae to better navigate future market dynamics. The ultimate goal is a stable housing finance ecosystem where homeownership is attainable and financially sound for a wide swath of the American populace.

Financial Assumptions

- ***Average Interest Rate (AVG IR):*** The interest rate listed at 0.54% is the average monthly rate. The annualized interest rate is calculated to be approximately 6.42% (0.54% * 12). This rate is essential for computing the interest income from the performing loans (True Negatives) and for understanding the lost interest income from loans incorrectly classified as high risk and not issued (False Positives).
- ***Number of Payments:*** The 120 payments indicate a loan term of 10 years, assuming monthly payments. This term is used to determine the total interest income a loan would generate over its life if it performs as expected without default, and it also helps assess the period during which a loan could potentially default (become a False Negative), affecting Fannie Mae financially.
- **Average Original Unpaid Principal Balance:** This is the average starting balance for the loans in the portfolio, which stands at $202,297.38. This figure is key to calculating the total principal that Fannie Mae is exposed to and could potentially lose in the case of defaults (False Negatives).
- **Cumulative Interest Received to Fannie Mae:** The amount of $72,361.02 represents the total interest payments received by Fannie Mae on these loans. It's an important figure as it provides an insight into the revenue generated from the interest on the issued loans.
- **Total Cost for Delinquent Loans:** The figure of $274,658.40 is the cost incurred from loans that have defaulted. This cost includes the unpaid principal balance that could not be recovered and any additional lost interest or costs associated with the default.

- **Overhead Costs:** In this analysis, we will not be including the overhead costs for simplicity sake, although they are noteworthy considerations since they are applicable in the real world.

AVG IR	0.54%		**6.42%/12**
# Payments	120		
AVG ORIGINAL_UNPAID_PRINCIPAL_BALANCE	$ 202,297.38		**AVERAGE(FM_2007Q1!E:E)**
Start PD	1		
End PD	120		
Type	0		
Cumulative Interest Received to Fannie Mae	$ 72,361.02		**ABS(CUMIPMT(B1,B2,B3,B4,B5,B6))**
Total Cost for Delinquent Loans	$ 274,658.40		**SUM(B7,B3)**

A breakdown the payoff matrix in this scenario would be the following:

- ***True Positive (TP):*** Loans that were predicted to default and did default. There's no direct financial impact from these loans because they were not purchased; however, opportunity cost or alternative investments that could have been made with the funds that would have gone to these loans can be considered.
    - Hypothetical Value: $0
- ***False Positive (FN):*** This would represent the loss Fannie Mae incurred by purchasing loans that were incorrectly predicted to be good but ended up defaulting. This figure would include both the principal that could not be recovered and the lost interest income from these loans.
    - Hypothetical Value: -$72,361.02
- ***True Negative (TN):*** These are the loans that were predicted to be good investments and indeed performed well. Fannie Mae would have purchased these loans, and the revenue from them would be reflected in the “Cumulative Interest Received to Fannie Mae,” which is $72,361.02. This represents the actual financial gain from interest on the performing loans.
    - Hypothetical Value: $72,361.02
- ***False Negative (FN):*** These are the defaulted loans that Fannie Mae did not anticipate would default and thus purchased. The cost here is represented by the 'Total Cost for Delinquent Loans' amounting to $274,658.40. This is the direct financial loss due to loans that became delinquent.
    - Hypothetical Value: -$274,658.40

Q1. Explore and pre-process data as needed. Provide a bullet list of pre-processing steps.

- ***Categorical Variables:*** Make sure CHANNEL, SELLER_NAME, FIRSTTIME_BUYER, LOAN_PURPOSE, PROPERTY_TYPE, OCCUPANCY, PROPERTY_STATE, ZIP_3, PRODUCT_TYPE, DELINQUENCY_STATUS are coded as categorical.
- ***Numerical Variables:*** Make sure ORIGINAL_INTEREST_RATE, ORIGINAL_UNPAID_PRINCIPAL_BALANCE, ORIGINAL_LOAN_TERM, LTV, CLTV, NUMBER_BORROWERS, DTI_RATIO, BORROWER_CREDIT_SCORE,

NUMBER_UNITS, MORTGAGE_INSURANCE_PER, and COBORROWER_CREDIT_SCORE are coded as numerical variables.

- ***Missing Values:*** There are no missing values in any of the columns, which simplifies the preprocessing step as we don't need to impute missing data.
- ***Variance of Features:*** Notable features with variance include BORROWER_CREDIT_SCORE, ORIGINAL_UNPAID_PRINCIPAL_BALANCE, and DTI_RATIO. Features like ORIGINATION_DAY and FIRST_PAYMENT_DAY show 0 variance, indicating no variability, so they should be excluded.
- ***High Cardinality Features:*** SELLER_NAME (13 unique values), PROPERTY_STATE (54 unique values), and ZIP_3 (892 unique values) exhibit high cardinality but are not removed due to the value they provide in the data analysis.
- ***Outliers:*** Detected in several numerical columns, such as MORTGAGE_INSURANCE_PER (30,124 outliers), LTV (9,883 outliers), and DTI_RATIO (5,696 outliers). In this analysis, I will not be modifying these outliers, as the analysis benefits from a more nuanced and comprehensive dataset. This enables a more accurate and meaningful interpretation of customer behavior.
- ***Exclude Features:*** Excluding features like LOAN_ID (due to its uniqueness and lack of predictive power), ZIP_3 as numerical (to avoid misinterpretation of geographical data), and PRODUCT TYPE (due to potential low variance, redundancy, or bias). We also have to drop original date features (ORIGINATION_DATE and FIRST_PAYMENT_DATE) as well as ORIGINATION_DATE Day of Month and FIRST_PAYMENT_DATE Day of Month since granular time-based features are included to avoid high cardinality and simplify models. These exclusions help in reducing overfitting, avoiding incorrect assumptions about ordinal relationships, minimizing noise, and ensuring the model focuses on variables with genuine predictive power and relevance.

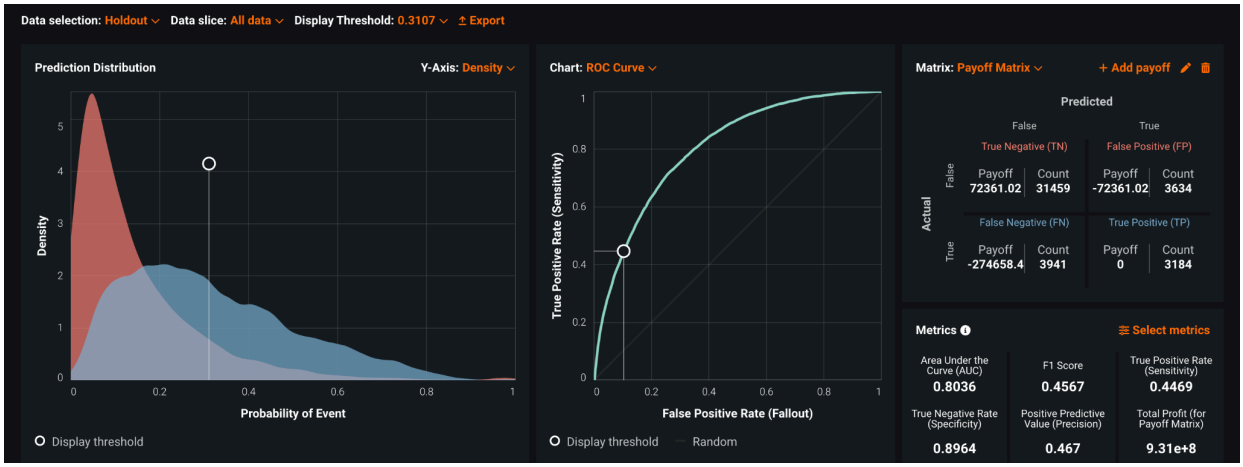
Feature Name	Data Quality	Index ↓	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
CHANNEL		2	Categorical	3	0					
SELLER_NAME		3	Categorical	13	0					
ORIGINAL_INTEREST_RATE		4	Numeric	296	0	6.25	0.38	6.25	2.67	9
ORIGINAL_UNPAID_PRINCIPAL_BALANCE		5	Numeric	605	0	202,297	96,926	188,000	6,000	802,000
ORIGINAL_LOAN_TERM		6	Numeric	58	0	360	1.47	360	301	360
ORIGINATION_DATE (Day of Week)		7	Categorical	7	0					
ORIGINATION_DATE (Month)		7	Categorical	12	0					
ORIGINATION_DATE (Year)		7	Numeric	8	0	2,006	0.51	2,006	1,999	2,007
FIRST_PAYMENT_DATE (Day of Week)		8	Categorical	7	0					
FIRST_PAYMENT_DATE (Month)		8	Categorical	12	0					
FIRST_PAYMENT_DATE (Year)		8	Numeric	8	0	2,007	0.25	2,007	1,999	2,007
LTV		9	Numeric	97	0	71.07	15.83	77	1	97
CLTV		10	Numeric	114	0	72.96	16.73	78	0	136
NUMBER_BORROWERS		11	Numeric	6	0	1.54	0.51	2	0	5
DTI_RATIO		12	Numeric	65	0	37.18	13.46	38	0	64
BORROWER_CREDIT_SCORE		13	Numeric	379	0	720	68.15	728	0	850

DTI_RATIO		12	Numeric	65	0	37.18	13.46	38	0	64
BORROWER_CREDIT_SCORE		13	Numeric	379	0	720	68.15	728	0	850
FIRSTTIME_BUYER		14	Categorical	3	0					
LOAN_PURPOSE		15	Categorical	3	0					
PROPERTY_TYPE		16	Categorical	5	0					
NUMBER_UNITS		17	Numeric	4	0	1.03	0.23	1	1	4
OCCUPANCY		18	Categorical	3	0					
PROPERTY_STATE		19	Categorical	54	0					
ZIP_3 (Categorical Int)		20	Categorical	892	0					
MORTGAGE_INSURANCE_PER		21	Numeric	17	0	3.31	8.45	0	0	40
[Few values] PRODUCT_TYPE		22	Categorical	1	0					
COBORROWER_CREDIT_SCORE		23	Numeric	347	0	314	364	0	0	837
DELINQUENCY_STATUS	TARGET	24	Boolean	2	0	0.17	0.37	0	0	1

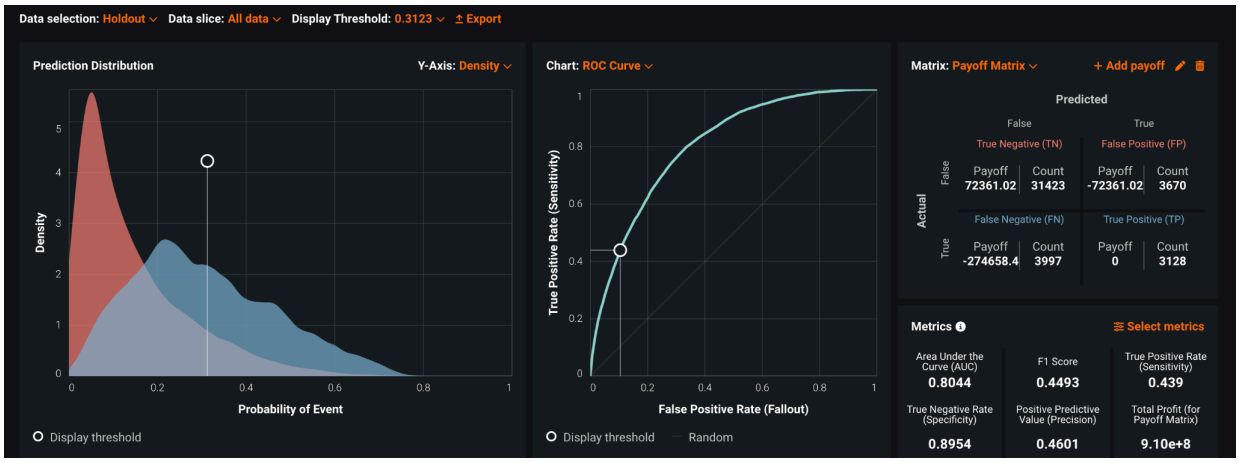
Q2. Make use of all the modeling techniques that you know to build models to forecast mortgage delinquency.

Nystroem Kernel SVM Classifier One-Hot Encoding | Missing Values Imputed | Standardize | Smooth Ridit Transform | Nystroem Kernel SVM Classifier M73 BP46	SallyMae_Preprocessed_Reduced 64.0 %	0.3657	0.3642	0.3621
Gradient Boosted Trees Classifier Ordinal encoding of categorical variables | Missing Values Imputed | Gradient Boosted Trees Classifier M12 BP35 REF SCORING CODE	SallyMae_Preprocessed_Reduced 64.0 %	0.3695	0.3689	0.3668
Logistic Regression One-Hot Encoding | Missing Values Imputed | Standardize | Logistic Regression M18 BP31 REF β_i SCORING CODE	SallyMae_Preprocessed_Reduced 64.0 %	0.3727	0.3720	0.3717
RandomForest Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Classifier (Gini) M66 BP44 SCORING CODE	SallyMae_Preprocessed_Reduced 64.0 %	0.3755	0.3759	0.3747

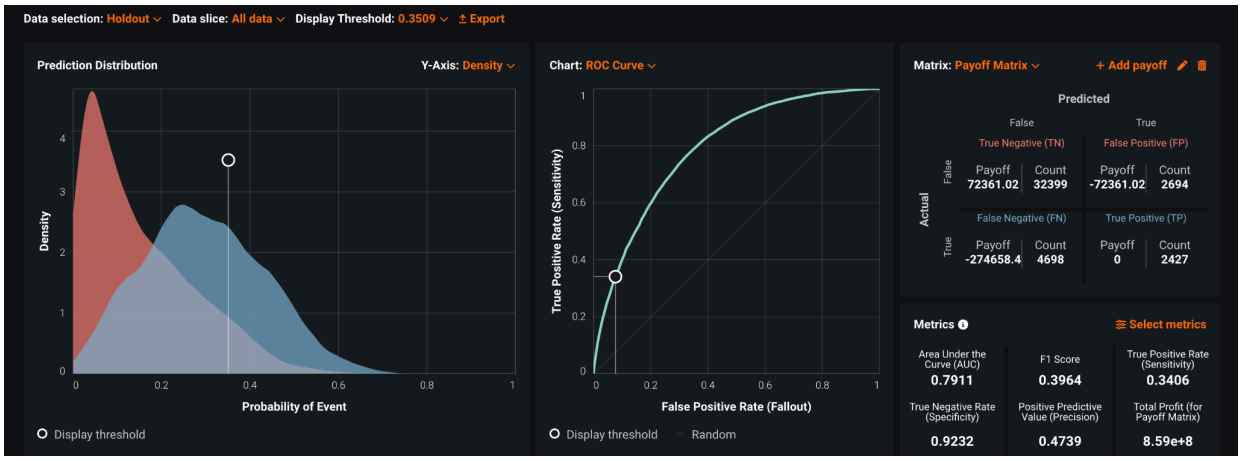
Logistic Regression: In the Fannie Mae dataset, logistic regression can be used to estimate the probability that a given mortgage will become delinquent. The model can incorporate various loan features such as loan-to-value ratio (LTV), credit scores, loan term, and interest rate as independent variables.



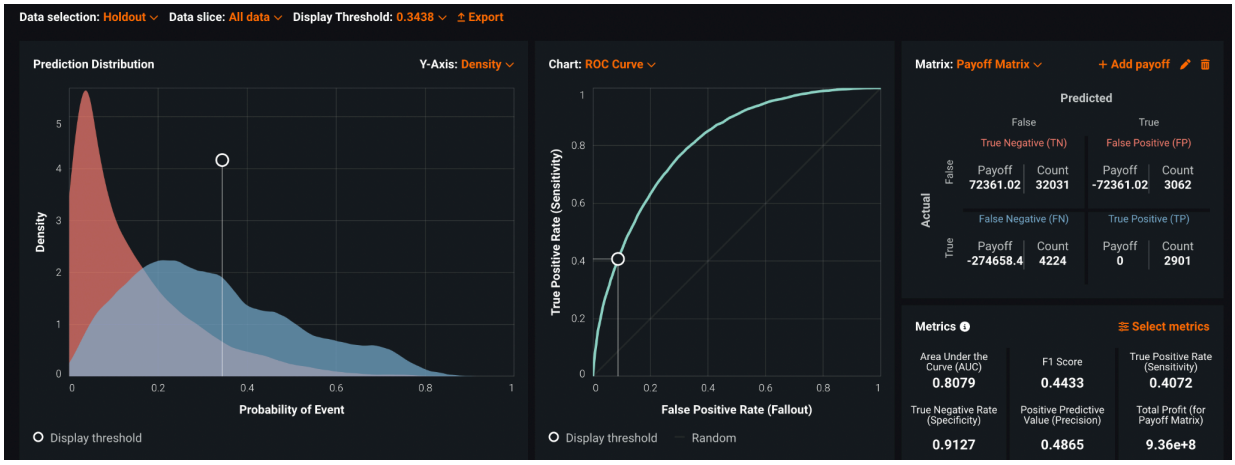
Gradient Boosted Trees Classifier: This model can handle the complex nonlinear relationships and interactions between features in the Fannie Mae dataset that may affect delinquency rates.



RandomForest Classifier (Gini): Random Forest can be used for its robustness and ability to model the delinquency risk without overfitting to the noise in the Fannie Mae dataset.



Nystroem Kernel SVM Classifier: The SVM classifier can be applied to capture complex relationships in the data. The Nystroem method is a technique to allow SVMs to scale to larger datasets, like that of Fannie Mae, by approximating the kernel function.



Q3. Show model performance metrics: Recall, Precision, Specificity, F1, ROC AUC, max payoff. What is the best metric to evaluate model performance?

	Logistic Regression	Gradient Boosted Trees Classifier	RandomForest Classifier (Gini)	Nystroem Kernel SVM Classifier
Recall	0.4469	0.439	0.3406	0.4072
Precision	0.467	0.4601	0.4739	0.4865
F1	0.4567	0.4493	0.3964	0.4433
Specificity	0.8964	0.8954	0.9232	0.9127
ROC AUC	0.8036	0.8044	0.7911	0.8079
Maximum Payoff	$931,000,000	$910,000,000	$859,000,000	$936,000,000
LogLoss	0.3727	0.3695	0.3755	0.3657
Cross Validation	0.3720	0.3689	0.3759	0.3642
Holdout	0.3717	0.3668	0.3747	0.3621
Threshold	0.3107	0.3123	0.3509	0.3438

Explaining the Metrics

- ***Recall:*** The proportion of actual positive cases correctly identified by the model. High recall means few false negatives.
- ***Precision:*** The proportion of positive identifications that were actually correct. High precision means few false positives.

- ***F1:*** The harmonic mean of precision and recall. It balances both to provide a single measure of a test's accuracy.
- ***Specificity:*** The proportion of actual negatives correctly identified. High specificity means few false positives.
- ***ROC AUC:*** The Area Under the Receiver Operating Characteristic curve. It measures the model's ability to distinguish between classes.
- ***Maximum Payoff:*** An estimate of the maximum financial benefit that could be achieved with the model's predictions.
- ***LogLoss:*** A measure of the accuracy of a classifier. It penalizes false classifications; lower values indicate better models.
- ***Cross Validation:*** A technique to assess how the results of a statistical analysis will generalize to an independent dataset. It helps in mitigating overfitting.
- ***Holdout:*** A portion of the dataset not used during model training, used later to test the model's performance to ensure it generalizes well.
- ***Threshold:*** The decision point between different classifications. Adjusting it affects the trade-off between recall and precision.

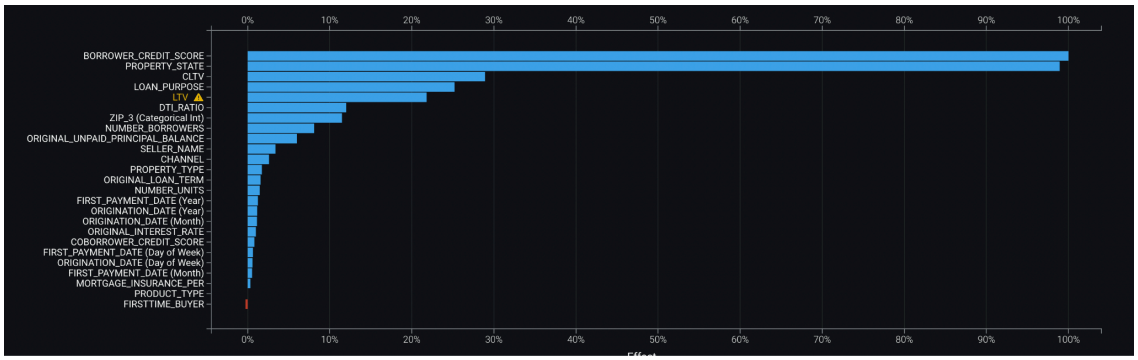
Best Metric

Given these models and metrics, if the primary goal is to maximize economic outcomes from predictive insights (which is often the case in financial settings like this), the Maximum Payoff would be the most relevant metric. It directly relates model performance to the financial benefits realized through correct predictions and effective interventions.

Best Model

- Given the context that the Maximum Payoff represents the potential profit and is the most relevant metric for evaluating the models, the Nystroem Kernel SVM Classifier is the best model. It has the highest Maximum Payoff at $936,000,000, indicating it could potentially provide the greatest financial benefit by accurately forecasting mortgage delinquency. This model not only promises the highest economic return but also exhibits strong performance across other evaluation metrics, including the highest ROC AUC score, which suggests it's particularly effective at distinguishing between classes (delinquent vs. non-delinquent loans). Therefore, based on the priority of maximizing financial returns through predictive accuracy, the Nystroem Kernel SVM Classifier stands out as the optimal choice.
- The Logistic Regression model, while offering substantial financial benefits, could not surpass the Nystroem Kernel SVM Classifier due to slightly lower precision in predicting loan delinquency and a marginally lower Maximum Payoff. Despite its competitive financial outcome, the Gradient Boosted Trees Classifier fell short of the Nystroem Kernel SVM Classifier because of its slightly lesser ability to maximize economic returns and a tad lower ROC AUC. Finally, the RandomForest Classifier fell behind the Nystroem Kernel SVM Classifier primarily due to its significantly lower Maximum Payoff and its reduced effectiveness in capturing all true delinquent loans, reflected in its lower recall rate.

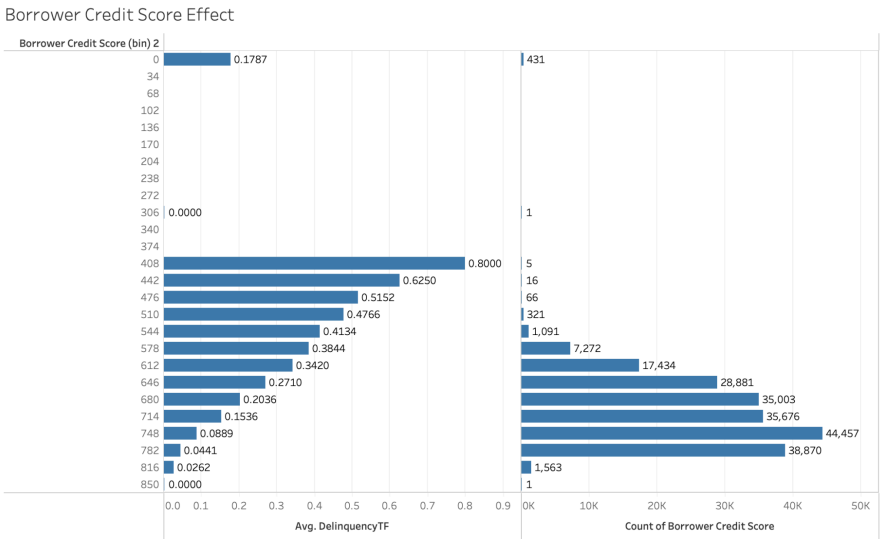
Q4. Visualize the effects of the top 4 predictors of mortgage defaults. Summarize the observed effects in 1-2 sentences for each predictor. Provide an actionable recommendation based on each observation. If the observation is not actionable, state so.



Using Datarobot, we are able to see the feature impact, excluding LOAN_ID, ZIP_3 (numerical), PRODUCT TYPE, ORIGINATION_DATE (Original), ORIGINATION_DATE (Day of the Month), FIRST_PAYMENT_DATE (Original), and FIRST_PAYMENT_DATE (Day of the Month). As such, for the Nystroem Kernel SVM Classifier model, the the most financially beneficial model with the highest Maximum Payoff, the top four predictors are:

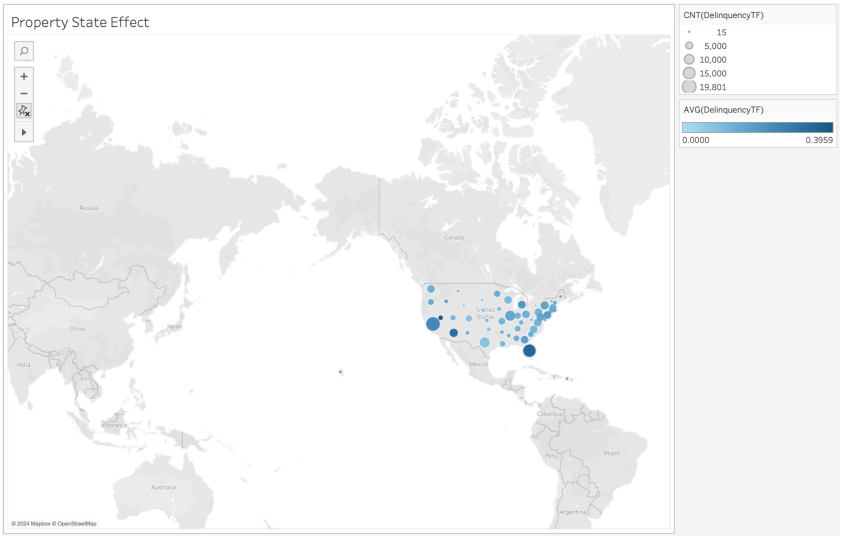
Borrower Credit Score:

- As the borrower's credit score increases, the average delinquency frequency decreases, highlighting a strong inverse relationship between credit scores and mortgage delinquency. Particularly, credit scores in the lowest bin show a significantly higher delinquency effect, whereas the highest scores correspond with virtually no delinquency, affirming the importance of the credit score as a predictor of loan performance.
- ***Recommendation:*** Borrowers with the lowest credit scores experience the highest delinquency rates, so prioritize loan approvals for borrowers with higher credit scores and consider implementing credit improvement programs for existing customers.



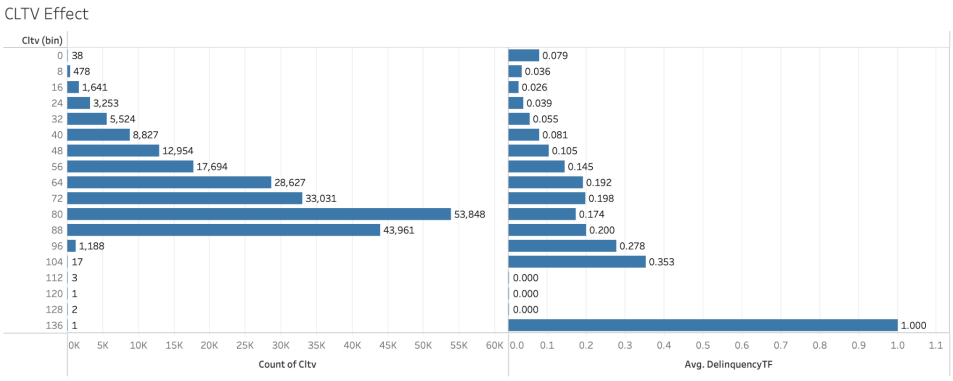
Property State:

- The state in which the property is located shows a substantial impact on default rates, indicating that certain states may have higher default risks, such as Florida and California, possibly due to economic variables.
- ***Recommendation:*** Adjust loan approval and pricing strategies state-wise to reflect the varying risk levels, and potentially develop state-specific foreclosure mitigation and borrower assistance programs.



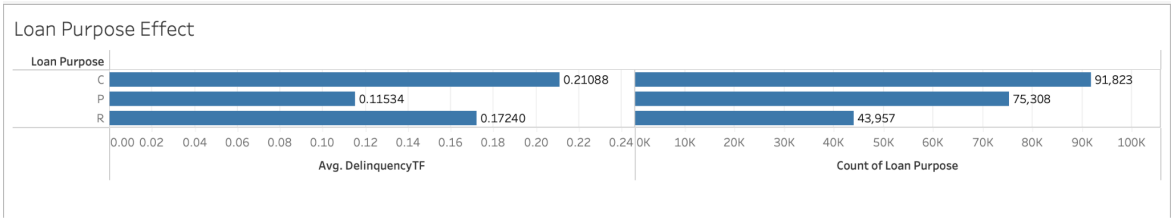
CLTV:

- There is a trend where higher Combined Loan-to-Value (CLTV) ratios are associated with increased delinquency rates, suggesting that as the amount borrowed approaches or exceeds the value of the property, the risk of default grows.
- ***Recommendation:*** Lenders like Fannie Mae should consider imposing more stringent lending criteria for loans with high CLTV ratios, potentially requiring additional credit enhancements such as mortgage insurance or higher interest rates to offset the increased risk. Additionally, they could provide targeted financial counseling for borrowers seeking high CLTV loans to ensure they understand the risks and obligations.



Loan Purpose:

- The chart indicates that the purpose of the loan is correlated with delinquency rates, with loans taken out for refinancing (R) having the lowest average delinquency, followed by loans for purchase (P), and loans for cash-out refinancing (C) having the highest delinquency rates.
- ***Recommendation:*** Develop stricter underwriting guidelines and possibly higher interest rates for cash-out refinancing loans to mitigate the higher risk of delinquency. For purchase and standard refinancing loans, which show lower delinquency rates, continue with current practices but monitor closely for any shifts in trends.



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

The models and metrics indicate that Fannie Mae did have information that could potentially have been used to predict mortgage defaults in Q1 2007. The data points such as loan-to-value ratio, borrower credit scores, loan term, interest rates, and loan purposes are all relevant predictors that, when properly modeled, can provide insights into the likelihood of a mortgage defaulting. The modeling techniques described, from logistic regression to machine learning methods like Random Forest and SVM classifiers, all have the capability to process these inputs and estimate default probabilities. Given the predictive performance of the Nystroem Kernel SVM Classifier, which showed the highest maximum payoff and strong ROC AUC, it's reasonable to conclude that models of similar sophistication, if they had been employed at that time with the same data, might have provided Fannie Mae with useful forecasts. However, the question of whether Fannie Mae could have accurately predicted defaults also depends on the quality and completeness of the data they had, as well as the advanced analytical capabilities available at the time. The tools and techniques for big data analytics have evolved considerably since 2007, and the predictive power of models has increased as a result.