SOHIL SHAH | ASSIGNMENT 5 DATA MINING MANAGERIAL

DELINQUENCY IN MORTAGAGE LOANS

Q1A

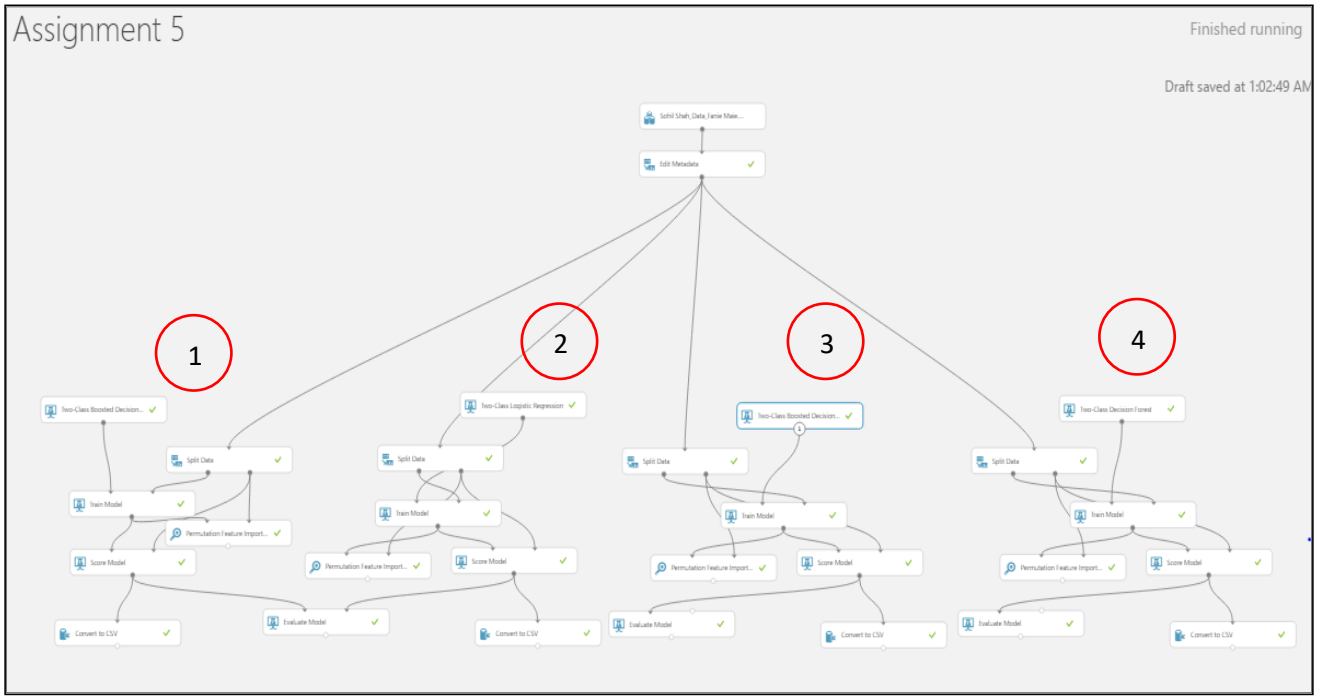


Figure 1: Work Space of Fannie Mae Delinquency Data (1) CART, (2) Logistic Regression, (3) Boosted Trees, (4) Decision Forest

Business Crux/ Overview of the Problem Statement:

The given data speaks about how Fannie Mae, a federally sanctioned corporation for promoting property ownership by buying up privately issued mortgages wrongly predicted during the housing mortgage demand during the 2007 Mortgage Crisis and was it aware of the situation. Through customer's Credit Score history and Delinquency Status we will further deep dive to investigate this scenario.

Steps for Data Cleaning & Refining:

1. Filtered the Q1 2007 data from the Origination Date column to focus as per the question
2. Recoded the BORROWER CREDIT SCORE column as those have Credit Score 0 will have the Credit Score of their Co – Borrower. In case Co-Borrower and Borrower both have a 0 Credit Score, we have removed them from the Analysis
3. Recoded and made new column BORROWER_CREDIT_SCORE_MODIFIED which has modified Credit Score by calculating Co – Borrower Credit Score for Borrower in case the Borrower does not have a Credit Score.

SOHIL SHAH | ASSIGNMENT 5 DATA MINING MANAGERIAL

4. Recoded a new column with range where depending on Scores we get Credit Ranges A, B, C and D where A is the best Credit Range.

   720+ → A

   690 - 720 → B

   630 - 690 → C

   350 - 630→ D

5. Recoded the Delinquency Status column which is out Target Variable from TRUE/ FALSE to 1/0 format.
6. Excluded the redundant columns like LOAN ID, ORIGINATION DATE, FIRST PAYMENT DATE, PRODUCT TYPE, BORROWER CREDIT SCORE, COBORROWER CREDIT SCORE (**NOTE:** As we now have BORROWER CREDIT SCORE RANGE) for the modeling purposes
7. Excluded some more columns like ORIGINAL LOAN TERM, ZIP 3, MORTGAGE INSURANCE PER, NUMBER UNITS and OCCUPANCY through Permutation Feature Importance
8. Defining the TARGET variable as Delinquency Status as we need to focus on knowing people are Delinquent or not

Q2

Metric	CART Model	Logistic Regression	Boosted Decision Trees	Decision Forests
FP	435	440	588	776
FN	3078	2968	2786	2888
Overall Error	0.168	0.163	0.161	0.175
Sensitivity	0.137	0.168	**0.219**	0.191
Specificity	0.975	0.975	0.966	0.955
F1	0.218	0.260	0.317	0.271
AUC	0.738	0.791	0.799	0.733

Considering the business case for Fannie Mae above, we have considered Delinquent status as 1 and Not Delinquent Status as 0. We see that most important part of the data would be in

SOHIL SHAH | ASSIGNMENT 5 DATA MINING MANAGERIAL

knowing Recall/ Sensitivity because we want to focus on records which were *predicted wrongly as not delinquent* or offering loan to people are actually delinquent but we identified as Not Delinquent and hence *False Negative* is the major factor and **RECALL** is the main metric.

The best model is *Two Class Boosted Decision Trees* (Here, I have used *50 Trees*) based on Recall value which is **21.9%** compared to other model values.

Q3

The main business scenario was to predict delinquency. If we introspect, we see that high credit score people were given more loan offers i.e. upto 100,000 people whereas people with lesser Credit score i.e. Class C were given less loan offers i.e. 35,000. But still Class C has more of Delinquents and % Delinquent is high among them.

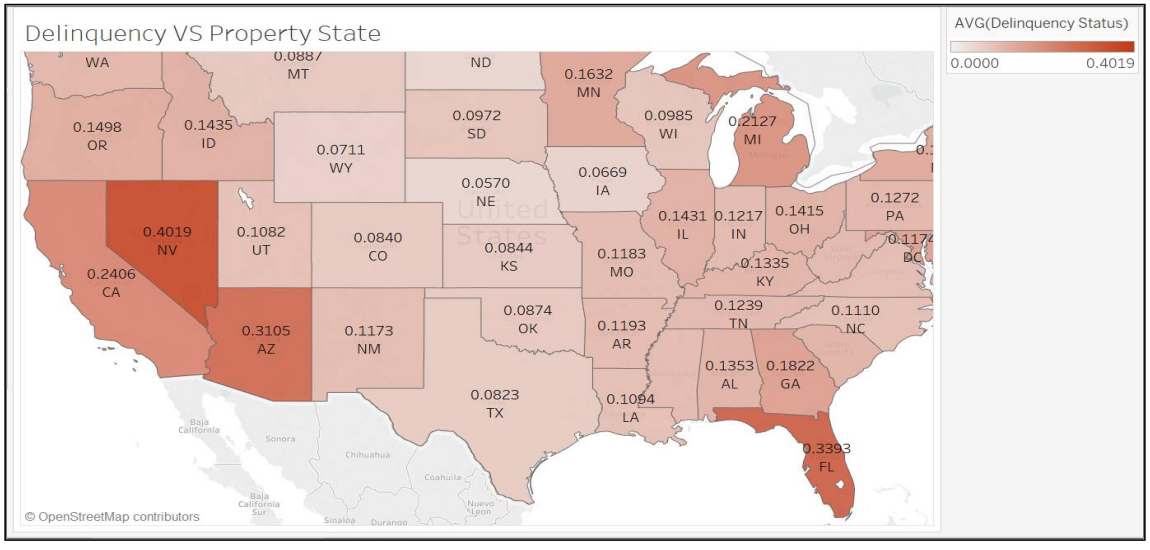
By inspecting the Scored dataset of the Two Class Boosted Decision Trees we see that Total False Negatives is 588, of which Class C i.e. people with lower credit score were having FN = 272. Thus, we see that around half of the wrong prediction for not delinquent people is among the major Delinquent Group. Thus, predictions were not much helpful if we compare the False Negatives values.

Thus, we see that through models also we get only 21.9 % Recall value in prediction which is not that much reliable. Hence, more data would have been better for them. Hence, I believe **they did not have enough information** for predicting the defaults.

Q4

As per the best model selected above, we have **Property State, Credit Score range** and **Loan Purpose** as the most Important Predictors.

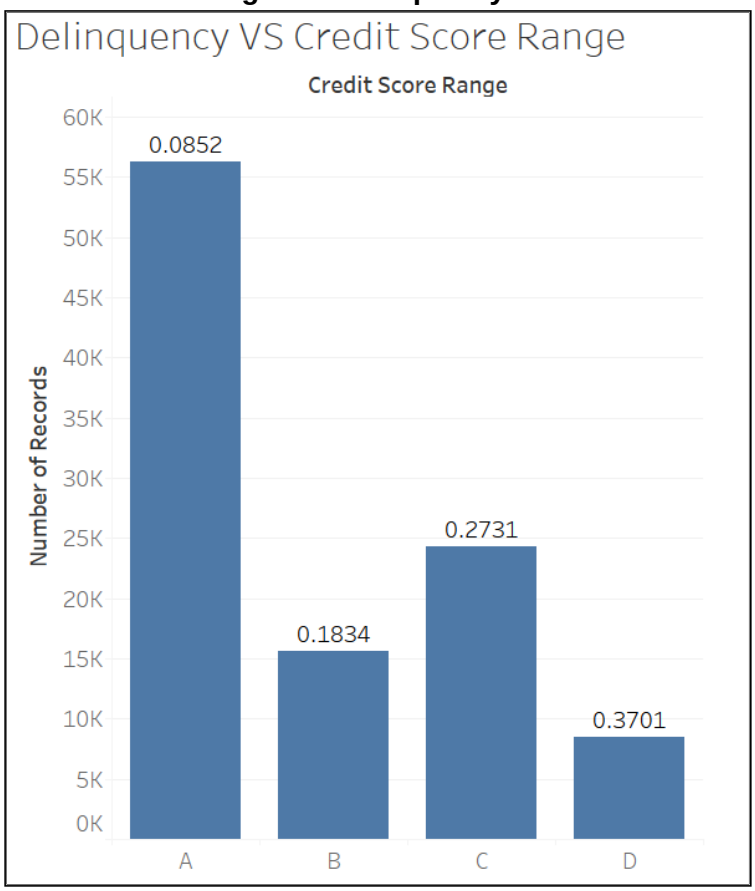
A. Property State VS Delinquency



SOHIL SHAH | ASSIGNMENT 5 DATA MINING MANAGERIAL

From the above, chart, we compare Delinquency across Property States. We learn that **Nevada (NV → 0.4019), Florida (FL → 0.3393)** and **Arizona (AZ → 0.3105)** are the main affected areas from **Delinquent Mortgage Loan Scenarios**. Also, one quick inference would be CA and FL had the maximum mortgage rates and housing prices, because many rich people stay here and they were given more loans.

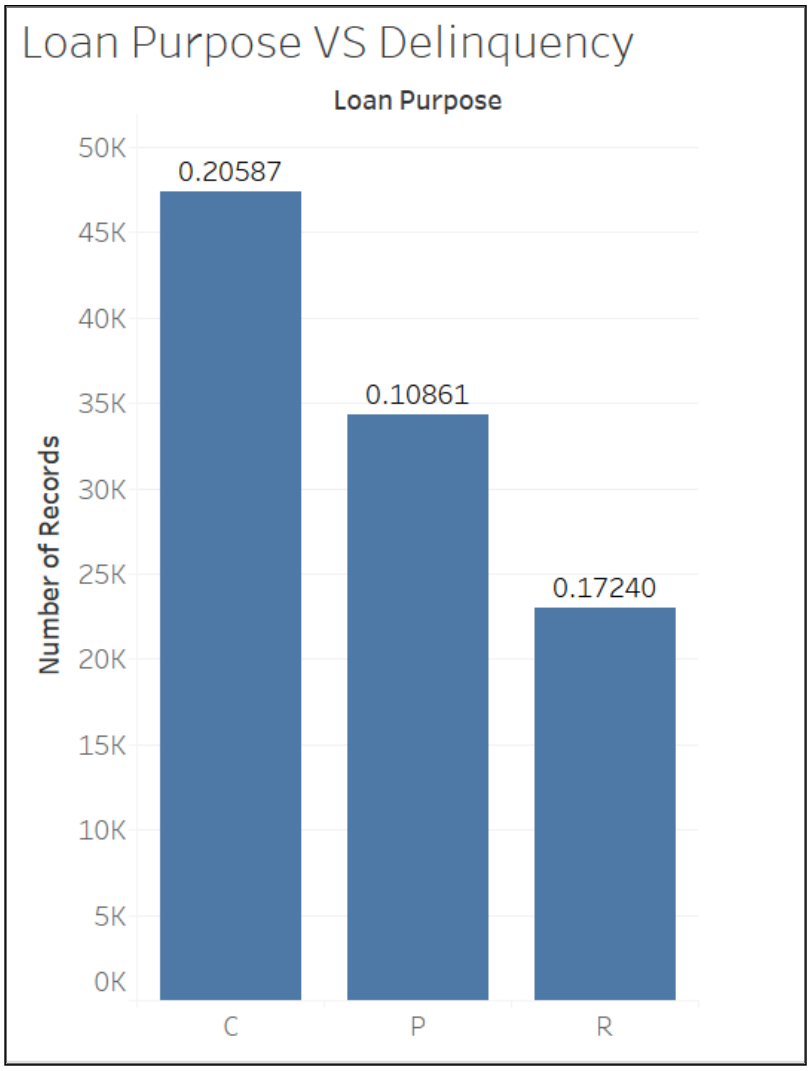
B. Credit Score Range VS Delinquency



From the above chart, we see based on Credit Score buckets and Delinquency Rate, there are around 56,000 loans given to **Class A** → High Credit Scorers and only **8.5%** of this was Delinquent. But, if we see **Class D → Poor Credit Scorers**, they have around 24,000 mortgage loans but **37% is delinquent** which is very much highlighter. Also, group C could have been a Profit class as they have higher interest rates but still there were more Delinquent cases.

SOHIL SHAH | ASSIGNMENT 5 DATA MINING MANAGERIAL

C. Loan Purpose VS Delinquency



As we can see from above visual, Delinquency across Loan Purpose categories. We see that **Cash-In Out [C]** has the highest mortgage loans and also highest delinquent rate of around **20.6 %**. Thus, there were many cash in-flows on the loan and more loans were taken at a lower interest rates and thus, this has been one of the contributing factors affecting Delinquency rates.