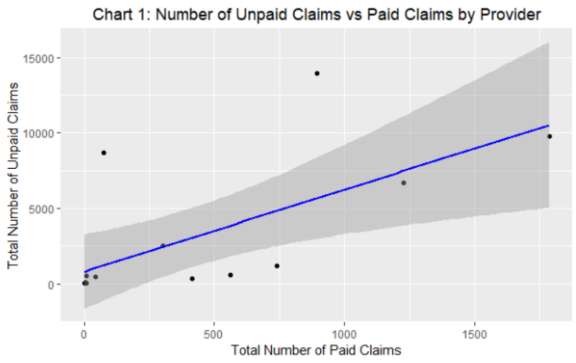
Author: Naveed Sharif
Data Source: Cogitativo
Topic: Data Science Challenge
Due Date: 10/28/2018

1. J-codes are procedure codes that start with the letter 'J'.
   a. Find the number of claim lines that have J-codes.
      - There are 51,029 claim lines with a J-code.
   b. How much was paid for J-codes to providers for 'in network' claims?
      - $2,417,221 was paid for J-codes to providers ‘in network’ claims.
   c. What are the top five J-codes based on the payment to providers?
      - The top 5 codes based on payment to provides are:
        (1) J1745 with a payment of $434,223
        (2) J0180 with a payment of $299,777
        (3) J9310 with a payment of $168,631
        (4) J3490 with a payment of $90,250
        (5) J1644 with a payment of $81,909

2. For the following exercises, determine the number of providers that were paid for at least one J-code.
   - There are 13 providers that were paid for at least one J-code

Use the J-code claims for these providers to complete the following exercises.

   a. Create a scatter plot that displays the number of unpaid for each provider versus the number of paid claims.



   b. What insights can you suggest from the graph?
   - The graph in chart 1 suggest that there is a positive linear relationship between the Total Number of Unpaid Claims and the Total Number of Paid Claims.
   - We can evaluate what the marginal impact of a Paid Claim is on Unpaid Claims by applying a simple linear regression;

$$Total\_Unpaid\_Claim = 789.4 + 5.5 Total\_Paid\_Claim_i + \varepsilon_i$$

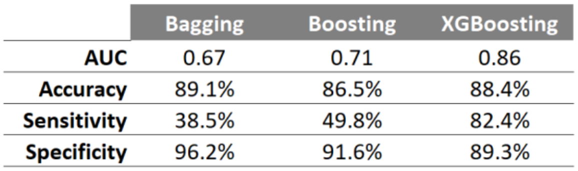
Suggesting that for every paid claim results in 5.5 unpaid claims. The Total Paid Claim is statistically significant with a pvalue of ~.01 and t-statistic of ~3.2.

Although this doesn't suggest causation because of the violation in the independence assumption; it does suggest correlation between Unpaid Claims and Paid Claims;

$$E(\varepsilon_i \mid Total\_Paid\_Claim_i) \neq 0$$

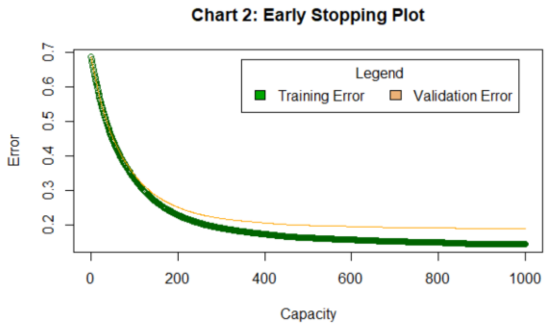
c. Based on the graph, is the behavior of any of the providers concerning? Explain.
   - Based on chart 1, there are two outlier Provider id's that are not following the linear patterns whereas the other Provider id's are.
   - Provider id FA0001387001's total unpaid paid claims is 7 times larger than what was predicted from the simple linear regression above (actual unpaid claims: 8,710 vs predicted unpaid claims: 1,193). While, Provider id FA0001389001's total unpaid claim is 2.5 times larger than what was predicted (actual unpaid claims: 13,947 vs predicted unpaid claims: 5,647).
   - They are concerning because we are underpredicting their total claims. This type of prediction (underpredicting) is much more impactful for the business than if the model was overpredicting (where prediction of unpaid claims is larger than actual unpaid claims).

3. Consider all claim lines with a J-code.
   a. What percentage of J-code claim lines were unpaid?
      - 88% of J-code claim lines were unpaid for (total J-code unpaid claims / total J-code claims).
   b. Create a model to predict when a J-code is unpaid. Explain why you choose the modeling approach.
      - I developed three different classification models, all of which are based on decision tree ensemble methods.
        (1) Bagging
        (2) Gradient Boosting
        (3) Xtreme Gradient Boosting (XGboosting)
      - I choose decision tree methods because (1) there are missing observations that would be dropped out if I used a regression algorithm such as a logistic regression, (2) decision tree ensemble models tend to do better in performing on unbalanced datasets.
      - This classification problem is an unbalanced dataset. 88% of the observations were unpaid.
   c. How accurate is your model at predicting unpaid claims?
      - The best performing model at predicting unpaid claims was based off the XGboosting algorithm. Since predicting unpaid claims is important for this business question, looking exclusively at the Accuracy is a misleading metric. Paying special attention to the Sensitivity Rate should be one of the focuses. Below, you will notice the XGboosting algorithm has the highest Sensitivity Rate (82.4%), highest AUC metric (0.86), and highest Accuracy (88.4%).

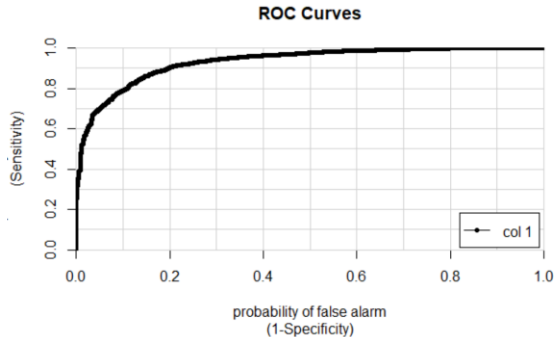


	Bagging	Boosting	XGBoosting
AUC	0.67	0.71	0.86
Accuracy	89.1%	86.5%	88.4%
Sensitivity	38.5%	49.8%	82.4%
Specificity	96.2%	91.6%	89.3%

- I used an Early Stopping plot to determine if the model was generalizing well or overfitting. Below, you can see that the Validation Error is low, but slightly higher than the Training Error. This tells me that the model has the ability to perform well on previously unobserved inputs, in other words the model generalizes well;

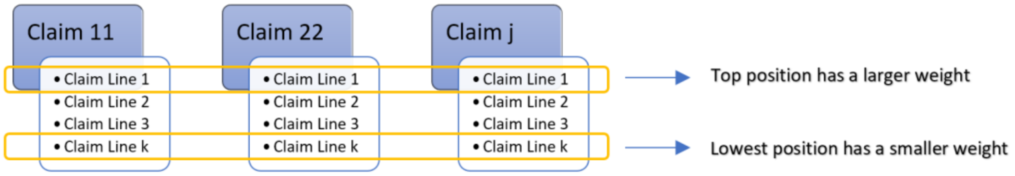


- Additionally, I used an ROC cure to determine the optimal threshold for the model;



d. What data attributes are predominately influencing the rate of non-payment?

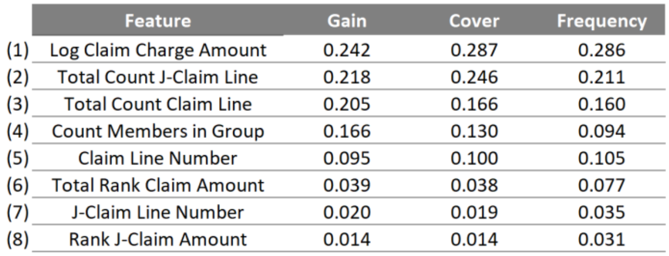
- In the past, I have worked on similar claims data business questions at Kaiser Permanente. What I have found was that the largest impact of predicting unpaid or delinquency claims are related to the size of the claim charge, the count of claim lines within a claim number, and the position of the claim line on the claim number (illustration in figure below);



- I featured engineered (FE) and included a series of variables I felt would impact an unpaid claim based on my experience and the available data in the claims dataset.
  (1) Log Claim Charge Amount (log to normalize the skewed data)
  (2) FE; Total Count of Claim Lines per J-Claim Number
  (3) FE; Total Count of Claim Lines per Claim Number

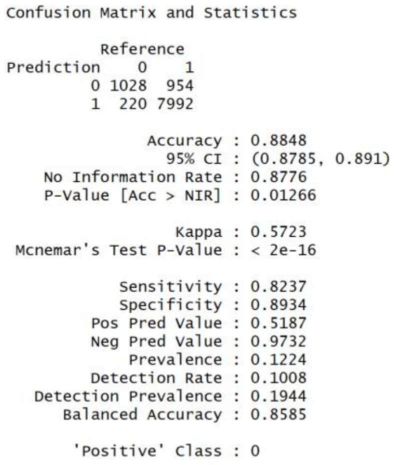
(4) FE; Total Members per Group
(5) Ranked Claim Line within Claim Number
(6) FE; Ranked Claim $ Amount by Claim Number
(7) FE; Ranked J-Claim Line within J Claim Number
(8) FE; Ranked Claim $ Amount by J-Claim Number

Feature Importance:



	Feature	Gain	Cover	Frequency
(1)	Log Claim Charge Amount	0.242	0.287	0.286
(2)	Total Count J-Claim Line	0.218	0.246	0.211
(3)	Total Count Claim Line	0.205	0.166	0.160
(4)	Count Members in Group	0.166	0.130	0.094
(5)	Claim Line Number	0.095	0.100	0.105
(6)	Total Rank Claim Amount	0.039	0.038	0.077
(7)	J-Claim Line Number	0.020	0.019	0.035
(8)	Rank J-Claim Amount	0.014	0.014	0.031

- There are several other features in the dataset, however the features I selected collectively performed best (selected features are in table above). Therefore, instead of creating a complex model with several features I built a flexible and simple model with the most important features at predicting an unpaid claim.
- Below are additional metrics from the XGboosting algorithm;



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1028  954
         1  220 7992

               Accuracy : 0.8848
                 95% CI : (0.8785, 0.891)
    No Information Rate : 0.8776
    P-Value [Acc > NIR] : 0.01266

                  Kappa : 0.5723
 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.8237
            Specificity : 0.8934
         Pos Pred Value : 0.5187
         Neg Pred Value : 0.9732
             Prevalence : 0.1224
         Detection Rate : 0.1008
   Detection Prevalence : 0.1944
      Balanced Accuracy : 0.8585

       'Positive' Class : 0
```