CLAIMS ASSIGNMENT

PREPARED BY NICK ARQUETTE (07/31/2020)

AGENDA

Challenge Overview

- Problem Statement

Data Overview

- General Procedure Questions
- Paid Versus Unpaid Info
- Provider Insights

Modeling

- Strategy
- Data Cleaning
- Pre-processing
- Evaluation and Review

Summary

- Feature Importance

Appendix (Auto-ML)

- Auto-ML (h2o)

ASSIGNMENT CHALLENGE

- What is correlated with Unpaid procedure claims?

DATA OVERVIEW


Procedure_Data_Report.html

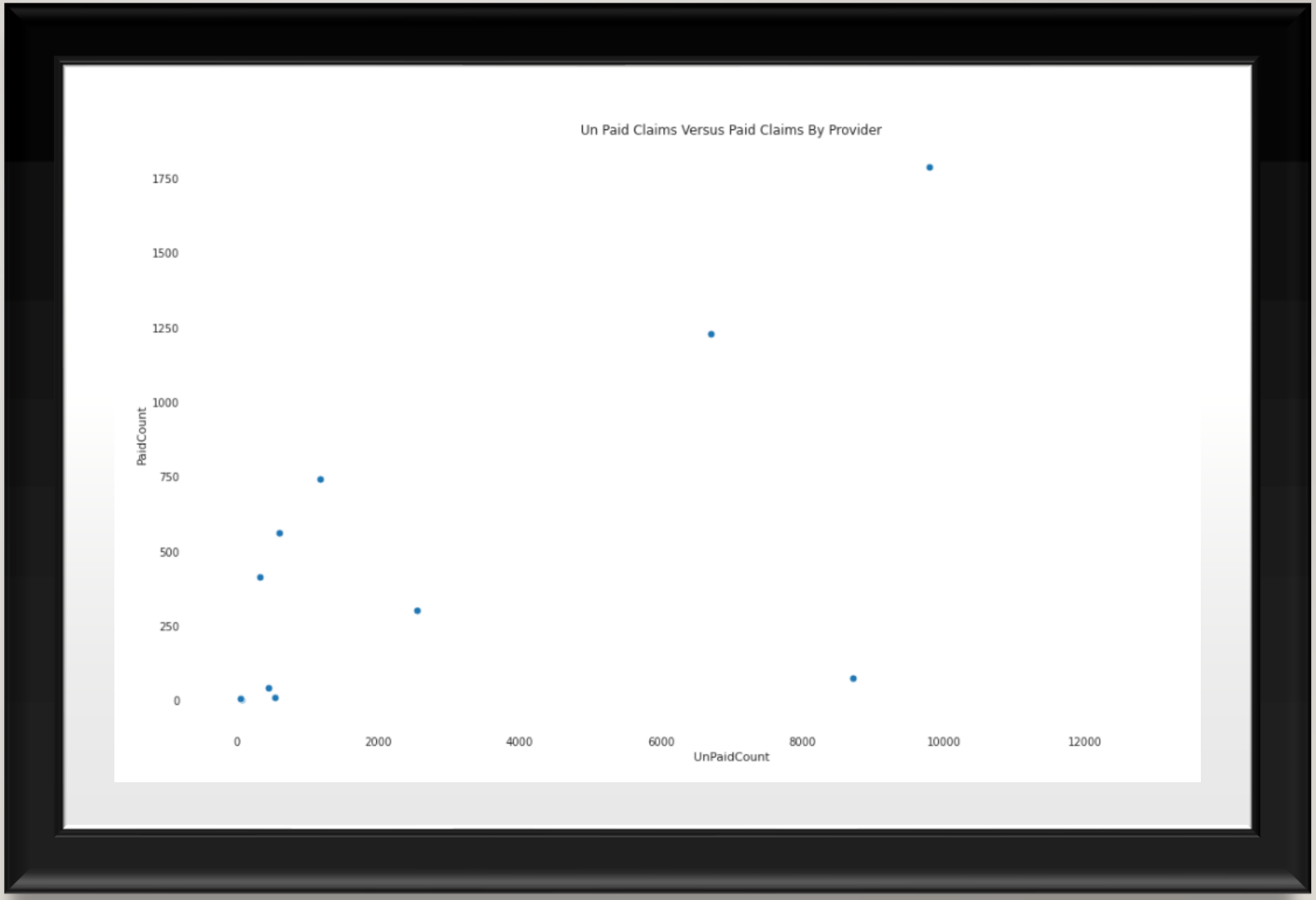
TOTAL CLAIMS

- 472,559 Total Rows
- 46,988 unique claims
- ~5 lines per claim (average)

PROCEDURE CLAIMS (CANCER)

- 10,691 unique claims **(Q1 Part A)**
- In-Network Procedure Payments
  - $ 2,471,220.96 **(Q1 Part B)**
- Top 5 Paid Procedures **(Q1 Part C)**
  - J2405, J2501, J7030, J1170, J1644
  - Cancer Related Procedures

PROVIDER UNPAID VS. PAID **(Q2 PART A)**



PROVIDER INSIGHTS **(Q2 PART B)**

Interpretation:

There is a positive correlation between the number of paid claims versus un-paid claims meaning that that are at least double the amount of unpaid claims per paid claim.

Concerns:

- Why are there some many unpaid claims?
- Why is different for one provider versus another?
- Loss in revenue for organization.

Questions:

What is the source of the problem (e.g. documentation, provider workflow, etc.)?

Are there any protocols in place to review this with management on a weekly basis?

What timeframe are we looking at?

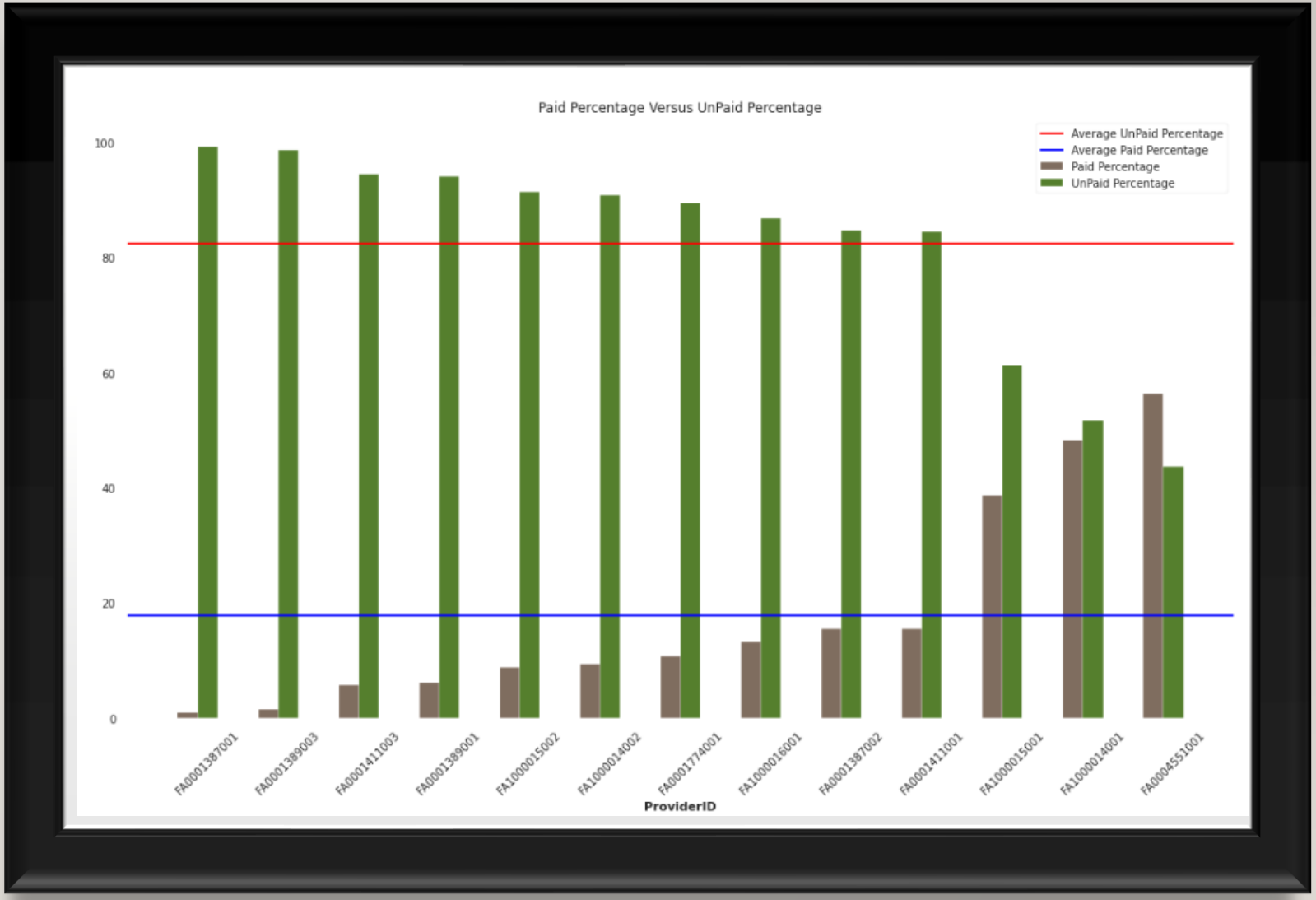
Are the numbers better for a different timeframe?

Should there be an unpaid percentage threshold?

AVERAGE UNPAID (82.31 %)

AVERAGE PAID (17.69 %)

(Q3 PART A)



AVERAGE UNPAID (82.31 %)

AVERAGE PAID (17.69 %)

(Q3 PART A)

ProviderID	PaidCount	UnPaidCount	Total Claims	Paid Percentage	UnPaid Percentage
FA0001389003	8	539	547	1.462523	98.537477
FA0001774001	302	2545	2847	10.607657	89.392343
FA0001387001	74	8710	8784	0.842441	99.157559
FA0001411001	1228	6703	7931	15.483546	84.516454
FA0001387002	1786	9799	11585	15.416487	84.583513
FA1000015002	43	449	492	8.739837	91.260163
FA0001411003	4	67	71	5.633803	94.366197
FA1000016001	7	46	53	13.207547	86.792453
FA1000015001	740	1170	1910	38.743455	61.256545
FA1000014002	5	49	54	9.259259	90.740741
FA0001389001	895	13947	14842	6.030185	93.969815
FA1000014001	561	601	1162	48.278830	51.721170
FA0004551001	415	322	737	56.309362	43.690638

MODELING **(Q3 PART B)**

- Predictive Output
  - Yes (Unpaid) or No (Paid)
- Modeling Algorithm
  - Supervised Classification Algorithms
- Base Model
  - Logistic Regression
- Competitor Models
  - Random Forest, Xgboost, Lgboost

Why?

- Appropriate to solve desired output
- Can handle different types of data (numeric and categorical)
- Can be visually interpreted by client
- Includes feature importance
- Addresses the problem of overfitting

DROPPED COLUMNS **(Q3 PART B)**

- Capitation.Index
- Claim.Current.Status
- Claim.Number
- Claim.Line.Number
- Claim.Pre.Price.Index
- Claim.Subscriber.Type
- Denial.Reason.Code
- In.Out.Of.Network

Why?

- Columns didn't add predictive value
  - Claim.Line.Number
  - Claim.Number
  - Claim.Subscriber.Type (all same value)
  - Member.ID
- Not enough data
  - Capitation.Index (many null)
  - Claim.Pre.Price.Index (many null)
  - In.Out.Of.Network (57 rows for Out of Network)
- Same meaning as UnPaid Claim
  - Denial Reason Code

DROPPED COLUMNS **(Q3 PART B)**

- Member.ID
- Place.Of.Service.Code
- Provider.Payment.Amount
- Subgroup.Index
- Subscriber.Index
- Subscriber.Payment.Amount

Why?

- Columns didn't add predictive value
  - Place.Of.Service.Code (all same value)
  - Subgroup.Index (most were 0)
  - Subscriber.Index (*same as member.id)
- Same meaning as UnPaid Claim
  - Provider.Payment.Amount

DATA BALANCING & CLEANING **(Q3 PART B)**

- Balancing
  - Under-sampled Paid Claims
    - Paid claims are the majority
- Cleaning
  - Claim.Charge.Amount
    - Remove outliers (negative numbers, and very large values)
    - Focus on data between the 5th and 95th percentile

DATA PRE-PROCESSING **(Q3 PART B)**

- Sklearn Pre-processing
  - Numeric Features
    - Median Imputation
    - Scaled (Standard Scaler)
  - Categorical Features
    - Constant Imputation ('missing value')
    - One Hot Encoding

Why?

- Models need all feature data to be in a numerical form
- Numerical features need to be put into a standard scale
- Missing data needs to be imputed so that information is not lost

MODELING EVALUATION **(Q3 PART B)**

- Fit Based Model
- Compare Base to Competitor Models using performance metrics (Mean Squared Error, Accuracy, AUC, etc.)
- Tune Models to determine best parameters
- Perform Cross-Validation using final parameters (e.g. 5-folds)

Why?

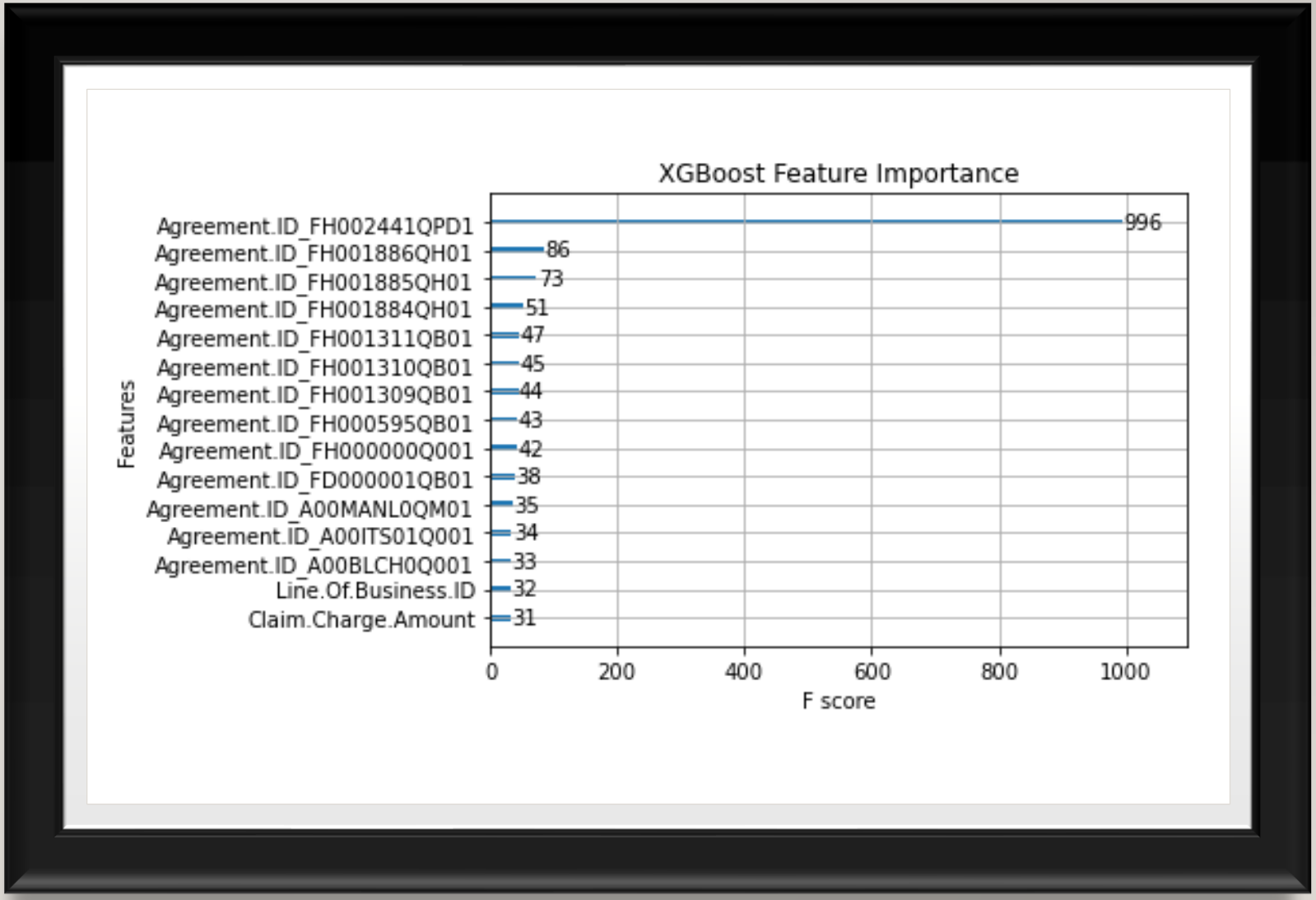
- Models parameter need to be tuned appropriately selected
- Cross Validation will help with estimate how the model will perform on unseen data

MODEL RESULTS (Q3 PART C)

- Best Model (Based On F1)
  - XGBoost
    - F1 = .9076
    - Recall = .9092
    - Precision .9061
    - Accuracy = .9074
    - MSE = .0925

Model	Mean SquaredError	Accuracy	Precision	Recall	F1
LogisticRegression	0.191851	0.808149	0.758179	0.904924	0.825077
RandomForest	0.170628	0.829372	0.784875	0.907470	0.841732
XGBoost	0.092569	0.907431	0.906091	0.909168	0.907627
LGBoost	0.186333	0.813667	0.762615	0.910866	0.830174

FEATURE IMPORTANCE FOR BEST MODEL (Q3 PART C)



FEATURE IMPORTANCE SUMMARY (Q3 PART C)

- Top 15 Feature Importance
  - Agreement.ID – Top 11
    - Assume this is an agreements between Practice and Insurance Company
    - Would want to understand the feature better
    - What agreement is this and why is it a problem
  - Line.Of.Business.ID –
    - Would like to understand the meaning behind the number and understand where the problem remains.
  - Group.Index
    - Would like to understand the meaning behind the number and understand where the problem remains.
  - Claim.Charge.Amount