Predicting Medical Claim Outcomes

A Study of Stony Brook Medicine Insurance Claims Distributed to Anthem Blue Cross Blue Shield

Casey Delaney

A paper presented for the degree of
Masters of Science, Analytics

Georgia Institute of Technology
College of Engineering
Atlanta, GA
July 2024

Abstract

The objective of this paper is to build an accurate classification model that can predict whether or not a medical claim sent by a Stony Brook Medicine provider to Anthem Blue Cross Blue Shield will be approved or denied, based on claim criteria such as coding used, amount charged, location of the patient, patient gender, and more. Another objective was to analyze claim outcome data, determining denial rates and other trends over time. The classification methods used include logistic regression and Naive Bayes. Features and parameters will be manipulated to find optimal outcomes. The goal of the research is to determine if claim filing can be improved in order to decrease denial rates.

From the research, it was determined that factors such as the frequency code, revenue codes, procedure codes, and insurance plans on medical claims can help predict whether a claim gets denied or not. While this may not benefit the patient, providers can determine how to code certain procedures to help ensure they get approved, as long as it is within reason. A Naive Bayes model proved to be the most effective in predicting whether a claim gets denied or not.

Table of Contents

1 Introduction

A healthcare claim denial can be a headache for both patients and healthcare providers. Claims denials from health insurance companies can occur for various reasons, some of which include a lack of prior authorization for the patient, the use of an out-of-network provider, or a lack of medical necessity. Some denials are easier to predict than others. However, many claims are denied unexpectedly, and services provided by healthcare institutions such as Stony Brook Medicine can end up partially paid or even unpaid, leaving either the patient, the provider, or sometimes both at a loss.

Currently, Stony Brook Medicine's Revenue Department does not have a method to predict claim results. An analytical approach of predicting claims denials would allow Stony Brook Medicine to set up for potential denials, while also aiding in preventing certain denials moving forward.

1.1 Purpose

The goal of this project is to help providers and decision makers within Stony Brook Medicine understand why claims are getting denied, and to determine if there are steps to be taken in the process of submitting of these claims to insurance companies that can ensure higher acceptance rates.

2 Patient Accounts

2.1 What is a Patient Account?

A patient account is a record that contains detailed information regarding a patient's medical charges and payments. Each patient account contains a wealth of information regarding the patient, including the patient's name, contact details, insurance information, medical diagnoses, procedures, medications, amounts charged for services, and much more. The accounts are used to process insurance claims and manage billing and collections of payments from the patient and/or insurance company.

An example of a fake patient account can be seen in Table 1 below, an actual account cannot be given for privacy reasons. It should be noted that there is much more information within each account, this is just a small sample.

Variable	**Value**	**Description**
pt_no	12345678	Patient Account Number
pt_name	Smith, John	Patient's name
pt_birth_date	1985-01-01	Patient's birth date
admit_date	2023-02-03	Patient's date of admission
dsch_date	2023-02-04	Patient's date of discharge
acct_type	IP	Inpatient or outpatient
balance	0.00	Balance on the account in USD ($)
tot_chgs	100.00	Charges for procedures in USD ($)
tot_pay_amt	100.00	Amount paid on the account thus far in USD ($)
payer_organization	Anthem	Insurance provider (payer)
product_class	Commercial	Product offered by payer

Table 1: Fake patient account example.

2.2 Account Criteria and Reasoning

It's important to understand what a patient account is because only certain accounts that fit a specific criteria will be used to complete this project. Only claims from accounts that met the following criteria were pulled:

- *Account Type:* Outpatient
- *Payer Organization:* Anthem Blue Cross Blue Shield
- *Product Class:* Commercial
- *Total Charges:* Greater than $0
- *Unitized Account:* Non-Unitized
  - A unitized account is used for a patient that has recurring visits. For example, a patient that is undergoing dialysis. They come in on a set schedule so each visit is given a unit number to distinguish them from one another. Using unitized accounts could dilute the data with many duplicates, as the patient information, charges, and other variables would all be the same.

Accounts with $0 in charges were useless, as this indicates a claim was never filed. Initially, that was the only filtering done. However, it was realized that the focus should be on non-unitized outpatient accounts utilizing Anthem as their insurance company. There are an abundance of these accounts, and this is often the criteria used for other Stony Brook Medicine process analyses. This would help determine the key indicators of the outcome of a claim, such as the procedures done, revenue codes used, insurance plans, patient zip code, etc. These could vary from payer to payer, complicating the results.

If the project yields positive results and provides key insights into denial data, the process will be done for other payers, products, and account types. However, it was important to establish this analysis as the baseline for future endeavors.

3 Data Extraction

All data was extracted from Stony Brook's servers using SQL. A view was created in order to replicate an extensive query that was created to pull all data needed. The view will allow experiments to easily be set up in the future.

3.1 How to Determine if a Claim is Approved or Denied?

It is common for a claim to receive more than one remit back from insurance companies. The reason for this is that if a remit is appealed, the following remit will yield the results of the appeal. This process can occur several times. It was important to just extract the first remit that was received back for each claim. This is because the goal of the model is to predict the initial feedback of a claim.

By comparing the total charges on a claim to the total amount paid by insurance, each claim was given an original classification of Approved, Denied, or Partially Denied. This was further classified into just Approved and Denied, where Partially Denied = Denied, as a claim that was not fully approved could not be classified as Approved.

3.2 Formatting each Claim to a Single Data Point

The last step before performing EDA and modeling was to get each claim into a single row, or data point. The issue was that each claim normally has multiple revenue codes (billing code used to specify services/procedures), procedure codes (code used to further specify procedures), and line amount (the amount charged for each line on a claim). A dynamic pivot was created in order to flatten out each claim and give each revenue code and procedure code its own column. For example, the revenue code 0510 became its own variable, labeled as "rev_cd_0510". Revenue code columns were filled with the amount charged for that revenue code on the entire claim, whereas procedure code columns were given a count of how many times the code appeared on the claim.

Due to there being almost 5,000 procedure codes, it wasn't feasible to create a SQL view as too many columns would be in the view; the limit was 1,024. So, only the top 100 procedure codes were taken and had columns created, all others were given a value of "Other", and a column was created for "Other".

3.3 Dataset Prior to EDA

Over 403,000 different claims, or data points, were pulled for analysis. These claims were all sent to Anthem between December 2017 and June 2024. A table of the variables, data types, and explanation of each variable can be found in Appendix A. Please note that the data type of several variables will be updated when performing EDA.

4 Exploratory Data Analysis

4.1 Data Cleaning Prior to EDA

The data was extracted into a CSV file, which was then pulled into R to conduct EDA and modeling. After this transfer, the data had to be manipulated further before exploration.

1. All NULL values were given a string value of "NULL" when they were pulled into R, rather than an actual NULL value of NA. So each "NULL" value was updated to NA.

2. Variables were removed that were unique to each claim. These include claim and patient specific IDs such as patient_account_number, pt_no, inst, medical_record_number, and inst_info_code. These were removed as they would not be useful moving forward, as new claims will have new values for each of these variables. They have no predictive power.

3. Certain string variables were pulled into R as numeric, so they had to be changed back to strings: facility_type_code, frequency_code, admission_type_code, patient_status_code.

4. Unit_no was removed as this was NA for all claims. Non-unitized accounts were the only accounts that met the criteria.

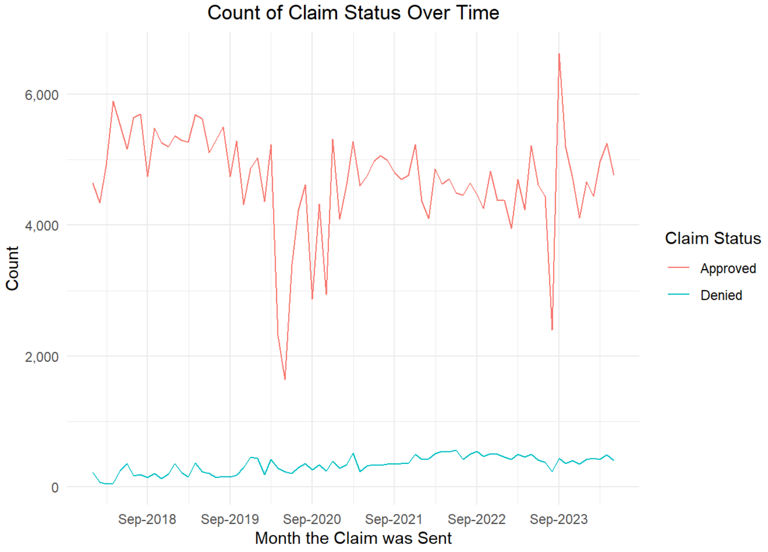
5. Rev_cd_[xxxx] variables were updated to have numerical values. Only the top 10 occurring of 121 revenue codes were kept in the data. Only the top 10 occurring procedure codes (px_cd_[xxxxx]) were kept as well.

6. Claims were removed from the data if the claim_status was anything other than Approved or Denied. The reasoning for this is that only claims with a definitive outcome should be analyzed. There were 13,658 claims removed with a claim_status of NA. This simply means that Anthem hasn't determined an approval or denial yet. There were 410 claims removed with a claim_status of Unknown. Unknown claims are claims that still under review, but Stony Brook has received a response. For example, Anthem may request to see further medical records and documentation for the patient before approving/denying a claim. Or, a claim that sent as a reversal for a previous payment that Stony Brook Medicine deemed as unwarranted.

7. Claims were removed if they were sent to Anthem on a Saturday or a Sunday. This indicates that the claims were sent in error and are likely duplicates, or don't provide enough information. These made up 0.15% of claims.

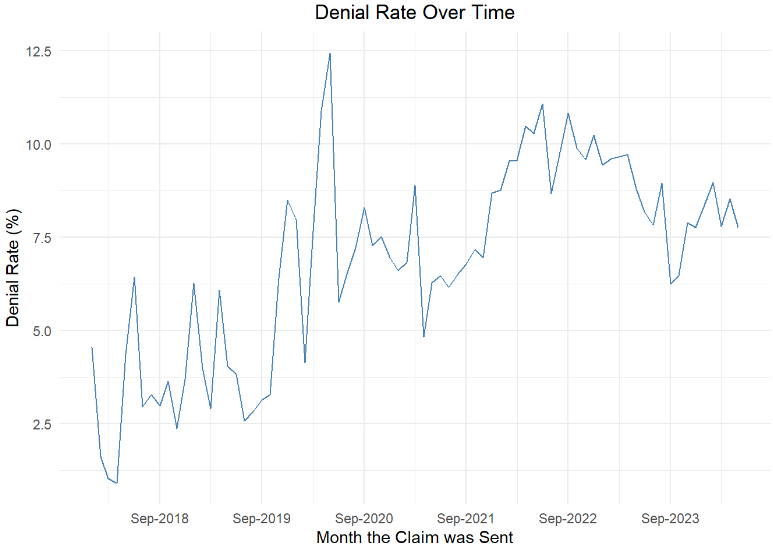
8. Lastly, claims sent in 2017 were removed, as they made up 0.29% of all claims.

4.2 Time Series Visuals



Graph 1: Count of Claim Status Over Time

Graph 1 illustrates the number of claims approved and denied over time. It shows that denials have steadily been increasing over time, whereas approvals have generally been decreasing. There are a few large spikes in approvals for claims that were sent in early 2020. This could be due to the initial stages of the COVID pandemic. What's interesting is that denials did not change much during this period, indicating the Anthem received less claims but increased their denial rate. There were two large spikes down and then up for approvals for claims sent around September 2023. A separate analysis would be required to determine the actual causes of the spikes in this visualization.

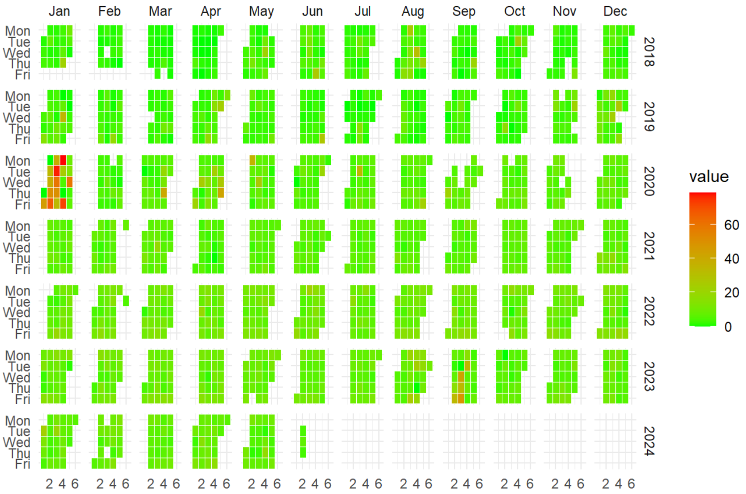


Graph 2: Claim Denial Rate Over Time

Graph 2 is a visualization of the denial rate for claims sent to Anthem over time. The denial rate was calculated by dividing the number of claims denied by the total number of claims sent to Anthem for each month claims were sent. It's easy to see that it has been increasing, which reinforces the data from Graph 1, where we saw a decrease in approvals and increase in denials.

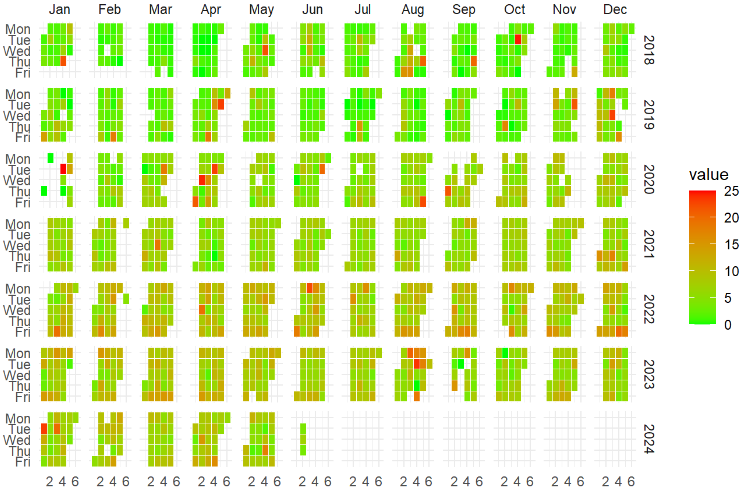
4.3 Denial Rate Heat Maps

It was shown how denial rate has trended over time, however it's important to break it down further to try and identify key time periods of higher rates that may not be seen as easily from Graphs 1 and 2.



Graph 3: Denial Rate Heat Map

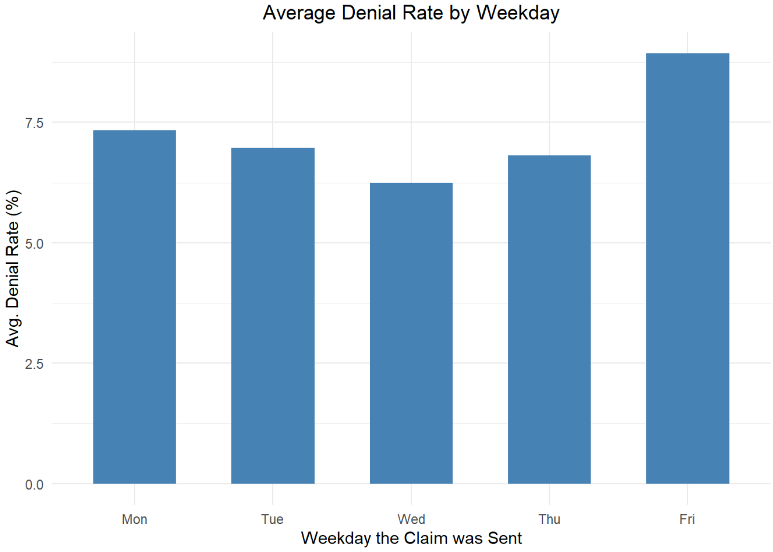
Graph 3 is a heat map that illustrates the denial rate for claims sent each day since 2018. The denial rate increases as the color of the box moves from green to red. It's evident that the majority of the highest denial rates are from claims sent in January of 2020, with a few darker orange boxes spread out throughout the data.



Graph 4: Denial Rate Heat Map - Denial Rates Less Than 25%

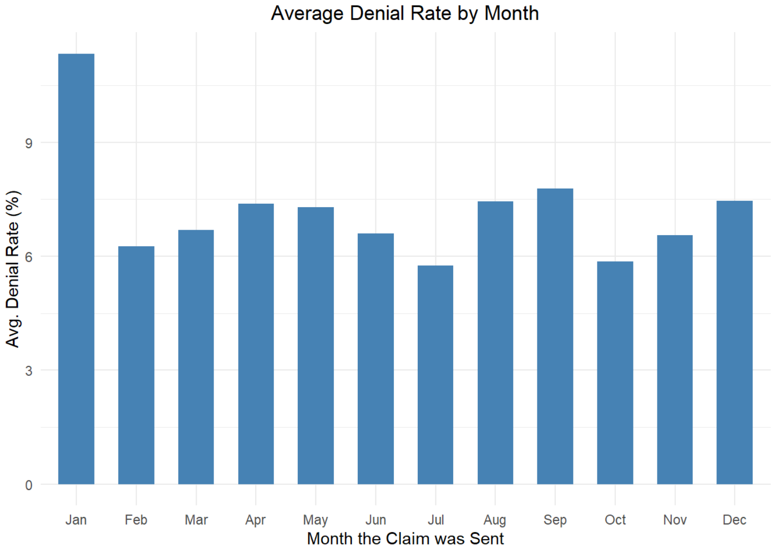
The majority of Graph 3 is green, making it difficult to see any other trends. After further analysis, it was determined that of the 1,630 total dates, only 1.96% of them are greater than 25%. Graph 4 shows a heat map with all denial rates that are less than 25%. Here, it's easier to see that there weren't certain months that increased overall denial rates other than January of 2020. As time moves on, the boxes get darker in color, indicating that there is a true trend of increasing denial rates.

4.4 Average Denial Rate Breakdowns by Time Period



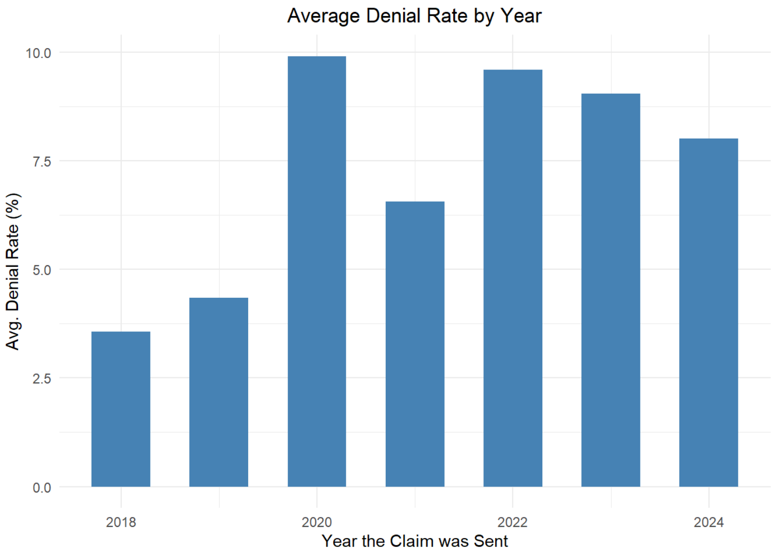
Graph 5: Average Denial Rate by Weekday

Graph 5 demonstrates that there's a v-shaped distribution of the average denial rate based on the weekday the claims were sent as the week goes on. It's interesting to note that Friday has a visibly high denial rate. It's possible that this is due to human interference, as people may spend less time looking at claims on Fridays, looking to get home. However, there's no way to accurately measure this based on the study.



Graph 6: Average Denial Rate by Month

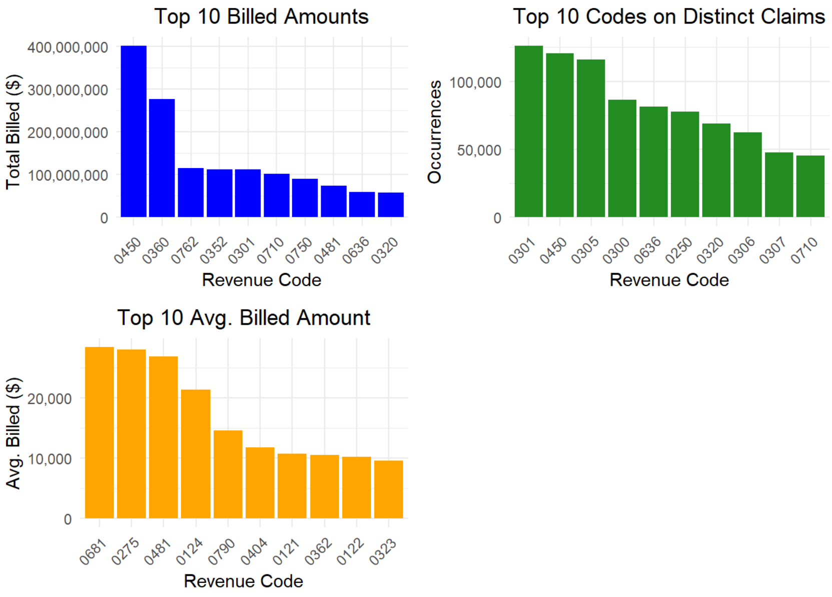
The highest average denial rate occurs for claims sent in January, by far. It also looks like that for the rest of the year, there's a pattern of increasing and then decreasing rates.



Graph 7: Average Denial Rate by Year

As seen in other visuals, 2020 has the highest denial rate of any year between 2018-2024. It must be noted that 2024 only contains data from half the year thus far, up until June. Rates have actually been decreasing since the second peak in 2022.

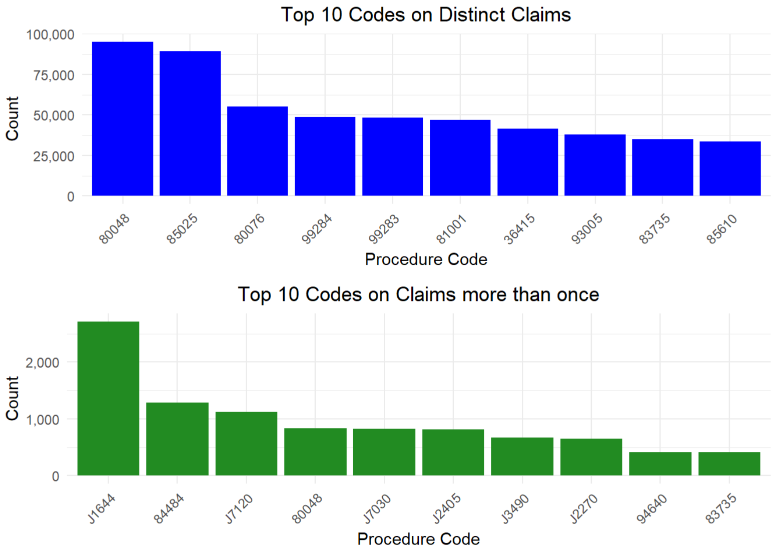
4.5 Revenue Codes



Graph 8: Revenue Code Metrics

Graph 8 is a breakdown of the most common and highest billed revenue codes found on the claims. Revenue code 0450, a code used for emergency room visits, was found on the second most claims and had the highest amount billed. The highest average amount billed per instance was revenue code 0650, a hospice service code. It's important to see the breakdown of the top revenue codes, as the modeling could demonstrate the effect that these codes have on the outcome of a claim.

4.6 Procedure Codes



Graph 9: Procedure Code Metrics

Graph 9 is a breakdown of the most common procedure codes found on claims. It's interesting to see that most of the procedure codes that appeared on claims multiple times were not the most frequent.

5 Modeling

5.1 Splitting the Data

The data eventually used for modeling was split into 70%-30% training and testing subsets. Random sampling was used to split, and equal proportions of Approved vs. Denied outcomes were in each subset. Specifically, 93% of claims in each subset were classified as Approved.

5.2 Logistic Regression

Logistic regression estimates the probability of an event occurring. In this case, the probability that a claim will be approved or denied. The binomial family was tested to measure the accuracy of predicting claim outcomes. Binomial was chosen because the outcome of a claim is binary; it is classified as either Approved or Denied.

5.2.1 Further Statistical Based Data Cleaning

Using all variables resulted in a highly inaccurate logistic regression model that took hours to run. Tens of thousands of dummy variables were created from the 254 original predictors, which led to most variables being insignificant and a heightened runtime. By identifying key variables that negatively impacted the model, it became increasingly effective and efficient with each change.

Uniform variables with only one unique value in the entire data set were removed. Variables that only have one unique value can cause errors when modeling, and won't be able to provide extra information about how each claim is different than another. Variables that had the same value for almost all claims were also removed, such as parent_name, info_code_qualifier, and billing_npi. These were removed from the data as only 80-90 of the over 403,000 claims had a varying value for each variable; essentially uniform throughout all claims. Subscriber_gender was removed, as over half of the claims had an NA value. Subscriber_zip was removed, with more than half of the claims having an NA value, and furthermore only 14 of the remaining 2,457 unique codes had a distribution of over 1%.

Operating_phys_id values were updated to 1 if the value was not NA, indicating that the patient had surgery. 80% of claims had an NA value prior to this binary value switch, with the rest of the claims having thousands of different IDs.

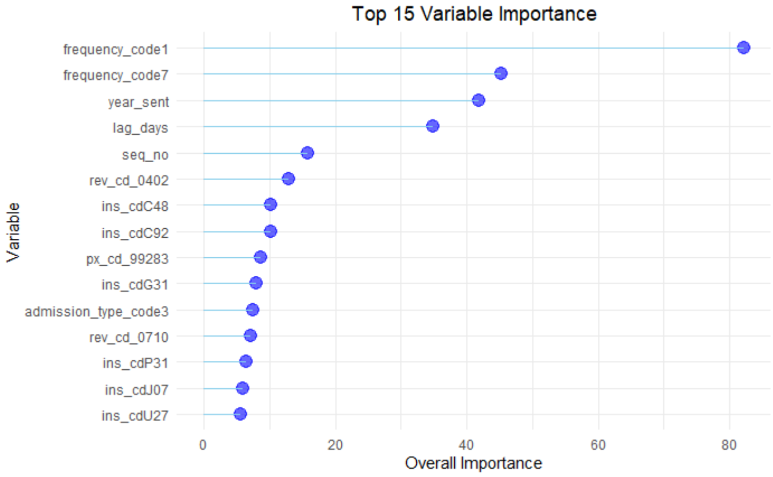
There were 121 rev_cd_[xxxx] variables present; this number was reduced to 10 to account for the top 10 most frequently occurring revenue codes. A similar process was done for px_cd_[xxxxx] variables, where only the top 10 most frequently occurring procedure codes were kept. These two steps were taken to simplify the model. Many of these codes were used infrequently, and provided no statistical significance.

Lastly, a lag variable was introduced. This lag accounted for the days between the date the patient was admitted into Stony Brook Medicine's provider (statement_from_date) and the date that the claim was sent to Anthem (file_creation_date).

5.2.2 Feature Selection

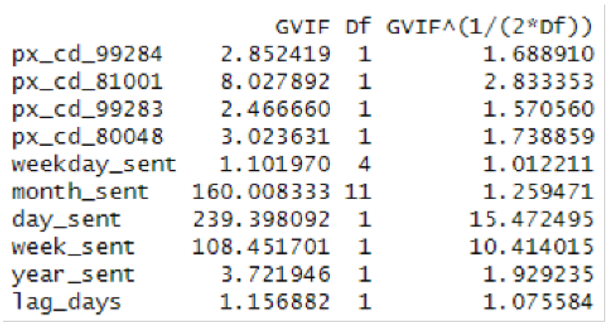
Through the data cleaning above, the model resulted in a root mean squared error (RMSE) of 0.20 and a prediction accuracy of 94.14%. At the surface, this appears to be an excellent outcome. However, it's important to note that the primary goal of the research is to predict claim denials, so it's also important to measure how accurate the model is with predicting a denial. Of 7,843 denials in the testing set, only 1,229 were predicted. This equates to just a 15.67% success rate. Perhaps further feature selection can help improve this.

Variable importance revealed that variables such as claim_filing_indicator_code and patient_status_code had a minimal effect on predicting the claim status, below 4.92e^{-3}. However, this does not mean that they do not have an effect. On the other hand, frequency_code, the year that the claim was sent, and the lag days had the highest correlation to predicting claim status.



Graph 10: Variable Importance

Next, variance inflation factor (VIF) was examined. The VIF is a quick measure of how much a variable is contributing to the standard error in the model [2]. A common rule of thumb is to remove variables with a VIF greater than 5 or 10. The GVIF is essentially the same as the VIF, and there were several values much larger than 10. However, this was mostly in categorical variables that were not scaled by their respective degrees of freedom. After scaling, almost all variables were below 3, however the day and week that the claim was sent were above 10. Note that Graph 11 below contains only 10 variables in the model.



```
                     GVIF Df GVIF^(1/(2*Df))
px_cd_99284      2.852419  1        1.688910
px_cd_81001      8.027892  1        2.833353
px_cd_99283      2.466660  1        1.570560
px_cd_80048      3.023631  1        1.738859
weekday_sent     1.101970  4        1.012211
month_sent     160.008333 11        1.259471
day_sent       239.398092  1       15.472495
week_sent      108.451701  1       10.414015
year_sent        3.721946  1        1.929235
lag_days         1.156882  1        1.075584
```

Graph 11: Variance Inflation Factor

Statistical significance via the model summary showed that while there were many variables that were statistically significant from alpha levels 0.001 through 0.1, many were not. However, it wouldn't be wise to pick out only certain dummy variables in groups such as ins_cd and month_sent to remove. There were others such as day_sent and all of the claim_filing_indicator_code dummy variables that would be better to choose for removal.

In the end, the chosen variables to remove were day_sent and claim_filing_indicator_code. Day_sent was removed for a few reasons. To start, it proved to have a high variance inflation factor. It was also statistically insignificant. Lastly, it had a near 1.0 correlation to week_sent, which makes sense because the day of the year that the claim was sent (numbered 1-365) would directly correspond to the week that it was sent (numbered 1-53). Highly correlated variables do not convey any extra information to the model. Week_sent was not removed because it was statistically significant. Claim_filing_indicator_code was removed because it had a very low variable importance and a p-value of approximately 0.99 for all dummy variables created.

5.2.3 Reduced Model Summary

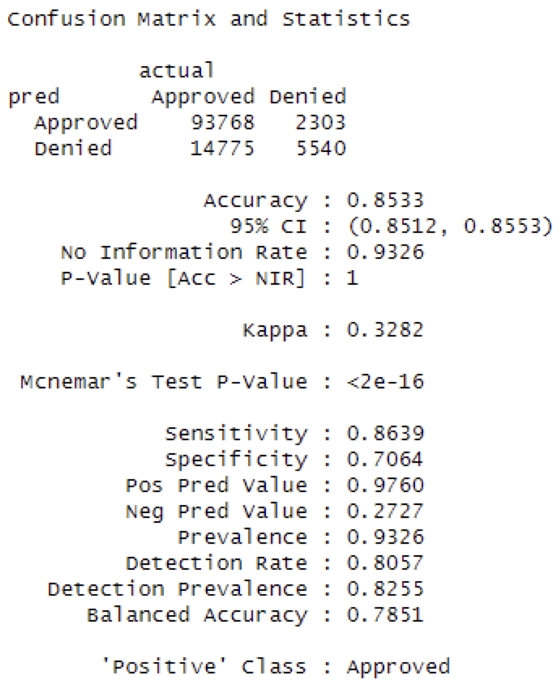
By reducing features, the RMSE did not change, staying at 0.20. In fact, even less denials were predicted. So, while logistic regression can produce a high accuracy, the primary goal of being able to predict denials is not being completed.

5.3 Naive Bayes

The Naive Bayes classifier calculates the conditional probability of a class based on prior knowledge gained during training. It is best used for larger data sets. However, a key note is that the Naive Bayes classifier assumes that each observation is independent of one another [4]. Naive Bayes was chosen because it would be interesting to see if there was a noticeable change in accuracy due to the fact that each observation would be deemed as independent.

5.3.1 Model Summary

The Naive Bayes model had a lower overall accuracy than both logistic regression models at 85.33%. However, it vastly outperformed when predicting denials. Naive Bayes yielded a 70.64% accuracy is predicting denials, which is a vast improvement over logistic regression. The table below outlines the model's confusion matrix:



```
Confusion Matrix and Statistics

          actual
pred       Approved Denied
  Approved    93768   2303
  Denied      14775   5540

               Accuracy : 0.8533
                 95% CI : (0.8512, 0.8553)
    No Information Rate : 0.9326
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3282

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

            Sensitivity : 0.8639
            Specificity : 0.7064
         Pos Pred Value : 0.9760
         Neg Pred Value : 0.2727
             Prevalence : 0.9326
         Detection Rate : 0.8057
   Detection Prevalence : 0.8255
      Balanced Accuracy : 0.7851

       'Positive' Class : Approved
```

Graph 12: Naive Bayes Performance

6 Conclusion

The final data set that was used to model the claims outcomes after removing, adding, and updating various variables consisted of 387,955 claims and 37 different predictors. It was unfortunate that the data didn't allow for a complete analysis of variables such as patient zip code and gender, as these could have been significant in predicting a claim's status.

To truly determine which classification method performed better, the goal of the research must be clearly defined. While logistic regression with full and reduced data sets performed better overall, the Naive Bayes method predicted claim denials much more effectively. So, what is better, incorrectly classifying approvals or denials?

A valuable metric healthcare institutions utilize in forecasting income and accounts receivables is the expected payments that they will receive on patient accounts and medical claims. If more claims are predicted to be approved, forecasts would look much better, as more payments will expected to be received. The issue with this is that the actual income would likely be lower. This would occur whether the logistic regression or Naive Bayes model were used. By utilizing Naive Bayes, actual income would appear to be less of a difference than the expected payment, leading to a more accurate forecast. So, to conclude, the Naive Bayes model despite being less accurate overall proved to be the better model in predicting medical claim denials.

6.1 Further Research

To start, further research will be done in fine tuning both models to improve on denials prediction rate. Different methods of classification such as support vector machines, K-Nearest Number, and decision trees will also be assessed.

An analysis of claims that all contained the gender or zip codes of patients could help in possibly identifying certain values that effected claim status.

Testing models on different institutions, product classes, and account types will likely yield different results for both the models used and others that will be created. It's certainly possible that it's easier to predict inpatient accounts vs. outpatient, or United Healthcare claims vs.

Anthem. There are a vast amount of opportunities for further research that can help Stony Brook Medicine improve their medical claim predictions and income forecasts.

7 Appendix

7.1 Appendix A - Dataset Prior to EDA and Modeling

Variable	**Type**	**x/y**	**Description**
claim_status	string	y	Approved or Denied
patient_account_number	string	x	unique identifier for each patient account that includes insurance plans and units
pt_no	string	x	patient account number
ins_cd	string	x	insurance plan
unit_no	integer	x	unit number (NULL)
seq_no	integer	x	internal sequence number
inst	string	x	claim number
file_creation_date	date	x	date the claim was sent to Anthem
billing_npi	string	x	national provider ID
claim_filing_indicator_code	string	x	tells Anthem whether the primary insurance is Medicare or another commercial payer
subscriber_zip	string	x	patient's zip code
subscriber_gender	string	x	patient's gender
charge_amount	float	x	amount charged on the entire claim
facility_type_code	string	x	first portion of the claim bill type
facility_code	string	x	Stony Brook Medicine database identifier
frequency_code	string	x	second portion of the claim bill type
statement_from_date	date	x	patient admit date
admission_type_code	string	x	type of admission; 1 = emergency, 2 = urgent, 3 = elective, etc.
admission_source_code	string	x	further classification of admission type cod
patient_status_code	string	x	code for a patient's discharge status
medical_record_number	string	x	patient's medical record number
attending_phys_id	string	x	attending physician ID
operating_phys_id	string	x	operating physician ID
rendering_provider_id	string	x	provider ID if different than attending/operating ID
parent_name	string	x	hospital name
into_info_code	string	x	line on a claim
info_code_description_short	string	x	fields on a claim; principal diagnosis, cause of injury code, reason for visit, etc.
info_code_qualifier	string	x	secondary code to identify the info code description
info_code	string	x	info code identifier

rev_cd_[xxxx]	float	x	a column for each revenue code on the claim, with the value being the total amount charged for that revenue code; otherwise NULL
px_cd_[xxxxx]	float	x	a column for each procedure code on the claim, with the value being the total count of the procedure code on the claim; otherwise NULL

Table 2: The dataset extracted from SQL.

8 References

[1] Bobbitt, Z. (2021). How to Fix: contrasts can be applied only to factors with 2 or more levels. *Statology* `https://www.statology.org/contrasts-applied-to-factors-with-2-or-more-levels/`

[2] Kelly, R., Li, T. (2024). Variance Inflation Factor (VIF). *Investopedia* `https://www.investopedia.com/terms/v/variance-inflation-factor.asp#:~:text=Variance%20inflation%20factors%20allow%20a,standard%20error%20in%20the%20regression`

[3] modelsummary: Data and Model Summaries in R. *Model Summaries* `https://modelsummary.com/vignettes/modelsummary.html`

[4] Ray, S. (2024). Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes. *Analytics Vidhya* `https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/`

[5] Yasar, Kinza. (2024). Logistic Regression. *Tech Target* `https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression`