Practicum Summary
DATS 599 - Master's Independent Study

"Predicting Outcomes of Medical Malpractice Claims"
Advised by Professor Alexander Olssen

Philip Situmorang

Contents

1 Background

The insurance industry offers value found in sharing risk. Most Americans readily see this through their own personal health insurance, which protects them from catastrophic financial loss due to medical events. Less obviously, having a personal health insurance also protects healthcare providers from financial loss, such as when the cost of a required medical treatment exceeds the individual's ability to pay. Insurance demand exists because risk sharing provides a more stable financial outcome for both individuals and hospitals, and such stability directly supports the operations of healthcare providers.

Medical malpractice insurance provides similar support to healthcare providers. Whereas health insurance limits an individual's financial loss due to healthcare costs, medical malpractice insurance limits a healthcare provider's loss due to medical malpractice claims. When practitioners obtain medical malpractice coverages, they effectively share their risks through the services of a risk pool manager, such as an insurance provider. This ability to share risk and limit possible loss assists providers in performing healthcare service to the public. For example, because of a malpractice coverage, a doctor who deems that a procedure's benefit far outweighs the risk may more confidently conduct such procedure, instead of avoiding it altogether out of fear of litigation and financial loss in case that the unlikely adverse outcome occurs.

Medical malpractice insurers must manage their pooled risk effectively to provide such support. Their value rests in the ability to pay when an insured must compensate the claimant for a malpractice act, and that ability to pay hinges on effective management of the risk pool. The tasks to effectively manage risk pool in the medical malpractice insurance realm are similar to tasks which belong to other sectors in the insurance industry. They include tasks such as determining the level of premium to charge and the level of reserves to keep.

With respect to premium, insurers theoretically balance charging a low enough rate to be competitive yet high enough to cover both the pool's expected loss and the insurers' service fees. With resepect to reserves, insurers balance between maintaining high enough liquid assets to quickly pay claims and yet not so high that not enough are invested in assets that, though not as liquid, will be better for the insurer's long-term growth and ability to pay claims. Sound analysis is key to effective risk management decisions involving these tasks, and it ultimately is key to nurturing a reliable support to providers.

Insurers generally analyze risk through two paradigms: probability and severity. Many tools are available to perform such analysis, and data is one of them. For the insurer, data provides insight into how cases involving covered risks with similar attributes have fared in the past. One source of such data is the insurer's internal data. For example, when one of a medical malpractice insurer's covered healthcare provider entered a litigation process, they can lean into their own experience to asses how cases with similar attributes have unfolded previously. Alternatively insurers may rely on an industry-wide aggregated data if one is available. For American medical malpractice insurers, such aggregate data is available through the U.S. Health Resources and Services Administration (HRSA)'s National Practitioner Data Bank (NPDB).

Making sense of the NPDB data to conduct a probability and severity analysis can be difficult given the sheer enormity of the data itself. This practicum seeks to simplify the analysis procedure, reducing the amount of time and difficulty to perform analysis that extracts insightful information regarding the probability and severity of a particular malpractice case. More specifically, the aim is to build a model which takes in as input attributes relating to a malpractice case and estimates two outputs: whether or not a particular malpractice case will result in some payment and the expected value of total payment given that the allegation results in a payment.

In this report we begin with analysis of the NPDB data, the basis of that model. This analysis

serves as the basis for selecting which variables of medical malpractice cases we will use in building the model. We will follow this analysis with a description of the methodology we used in building the two components of our model. And finally, we summarize the report with a brief conclusion which detail findings and limitations.

2 Exploratory Data Analyis

2.1 Overview

The National Practitioner Data Bank dataset contains 1,557,701 rows of data as of April 2021. Each row describes 54 variables relating to a particular medical malpractice case in the United States and its territories since 1990. Out of these, 14.4% (223,926 rows) are recorded as resulting in payment to the claimant. The remaining cases which do not result in a payment are labeled as "adverse action."

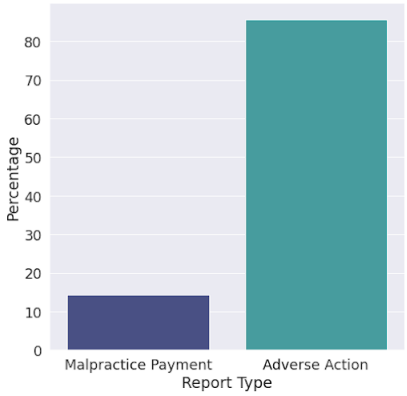


Figure 1: Comparison of number of cases which result in payment and those which do not

Among the above malpractice payment cases, the vast majority report payment of less than $1,000,000, unadjusted for inflation. This information is reflected specifically in the TOTALPMT variable column of the dataset. When adjusted for to Consumer Price Index (CPI) levels of April 2021, the percentage of cases resulting in payments of less than $1,000,000 drops from 96% to just over 90%. Close to 10% are now in the above $1,000,000 payment range, an increase from around 4% previously. This inflationary adjustment is a key step in developing a response variable for the severity estimation component of the model.

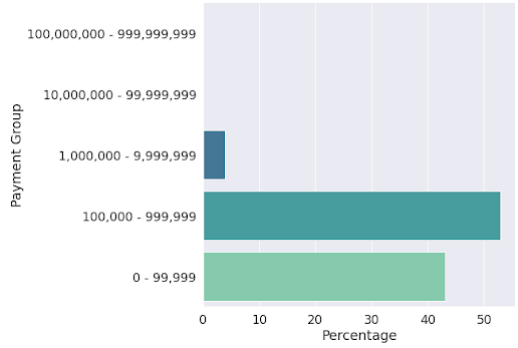


Figure 2: A historical overview of the severity of medical malpractice payments in the U.S.

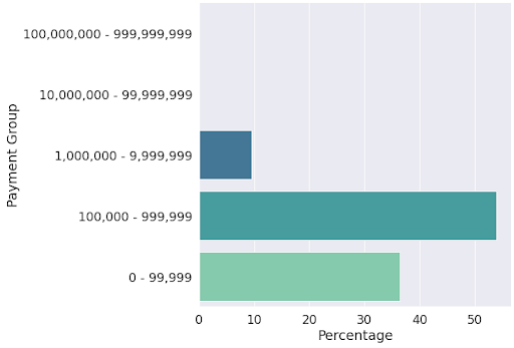


Figure 3: Medical malpractice payments adjusted to 2021 CPI levels

2.2 Defining Response Variables

The model's first component seeks to estimate the probability that a given case results in a payment. This objective requires a true or false response variable: true if the particular case row results in a payment and false if otherwise. We can simply derive this from the dataset's TOTALPMT variable, creating a new column where "0" indicates null or zero payment and "1" indicates a payment of greater than zero. It is key to note that we do not aim to build a classifier model. As the vast majority of malpractice cases result in no payment to the claimant, a classifier would almost always predict that a case will not result in a payment. Here a probability estimation is a more meaningful output, and thus a regression model estimating such probability is preferrable.

The model's second component seeks to estimate the severity of a payment *given* that a case results in a payment. That is, we estimate not the pure expected severity considering all cases but a conditional expected severity based only on cases which results in a payment. Similar as before, a conditional expected severity is more meaningful output than a pure expected severity estimation. Given that the first component already captures the probability of a case resulting in a payment, a more prominent question which follows is how severe would such case be *if* some payment is awarded to the claimant.

For the second component we use as response variable the payment listed in the TOTALPMT column *adjusted* for inflation. To perform the adjustment we obtain CPI data from the Bureau of Labor and Statistics page and select CPI as of June in each year since 1990, the earliest year found in the dataset. We multiply each row in the TOTALMPT column with an inflation factor: the CPI as of April 2021 divided by CPI as of June of the year associated with that particluar case. As the NPDB dataset specifies only the year of case, we arbitrarily select June CPI levels as the basis of adjustment.

2.3 Feature Selection

Out of the fifty-four variables available from the dataset we select the following as the final five variables from which to build the model:

1. LICNSTAT - The licensing state of the practitioner involved in the case
2. PRACTAGE - The practitioner's age group

3. LICNFELD - The practitioner's license field (i.e. a Registered Nurse or Osteopathic Physician)

4. ALEGATN1 - The primary allegation type (i.e failure to diagnose and improper treatment)

5. OUTCOME - The outcome of the case (minor permanent injury, death, etc.)

Multiple criterias serve as the basis to select these features. Below are the criterias which was used to select such features.

Statistical significance. We desire our variables to have statistical significance, which means we seek to select variables which can mathematically be substantiated to be "predictive." The selected features above are solely categorical. As we have a categorical response variable for the model's probability component, we conduct a Chi-square test to estimate significance of each variable. For the model's severity component and its continuous response variable, we conduct a partial F-test to estimate significance. We also test the assumptions which underlie the two significance methods.

Causal inference. Statistical tests may conclude that a variable is statistically significant, but at times a causal relationship can be difficult to infer between such variable and the response variable. Thus, the features above were selected not only on the basis of statistical significance but also based on the intuitiveness in how they cause the probability or severity of a medical malpractice payment to vary. As we intend to build a marketable model, such intuitiveness is key to build confidence in the model. For example, one can readily see that the selected variable of practitioner's age group correlates with experience. Subsequently, one can infer with relative ease a causal relationship between experience level and say, the probability of a practitioner performing an error during an operation.

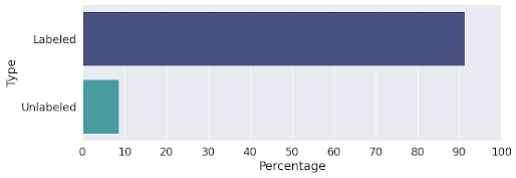
Ethics. Ethical considerations in selecting variables revolve primarily around patient privacy. The inclusion of patient-related variable may lead users of the model to inquire patient data which may lead to cases of privacy violation. We excluded patient-related features in our selection procedure specifically for this consideration.

2.3.1 Feature 1: Licensing State (LICNSTAT)

The selection of Licensing State as feature is based on the possible differences in licensing or regulatory standards that may impact the probability or severity of a malpractice payment. For example, with respect to licensing standards, states with higher licensing standards may indicate lower probability and severity. With respect to regulatory standards, states may enforce different medical malpractice laws which may impact how likely a judgment be awarded to the claimant or how much could be paid in a given malpractice condition. While it is possible for practitioners to practice outside their licensing state, the NPDB dataset indicates that the vast majority practices within their licensing states.

I. Feature Overview

Over 90% of the entire dataset labels the licensing state of the practitioner associated with the case.



The LICNSTAT feature has 57 categories, which include 50 U.S. States and 7 U.S. Territories such as the Virgin Islands and Puerto Rico. Texas, California, and New York are the three largest licensing states in the dataset.

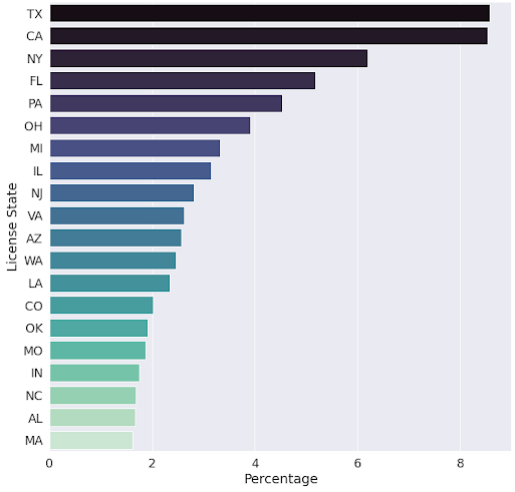


Figure 4: top 20 licensing states represented in the datase

II. Statistical Significance

To determine statistical significance, we conduct both Chi-Squared and F test for both the probability and severity components of the model. We will also try to estimate significance visually. We do so by looking at how probability and severity vary accross each licensing state. The response variable should vary as license state varies, if the variable is a statistically significant variable. We will see that this is indeed the case.

II.I. Chi-Square Test

The Chi-Square test between the categorical feature LICNSTAT and response variable PMT results in effectively zero p-value. This means we cannot reject the null hypothesis that there is no significance between the two variables. Note that PMT is a categorical response variable which indicates whether or not a particular case results in a payment or otherwise ("1" indicates true and "0" indicates false).

R Code:
licnstat.data = table(df$LICNSTAT, df$PMT)
chisq.test(licnstat.data)

Output:
Pearson's Chi-squared test
data: licnstat.data
X-squared = 95813, df = 58, p-value < 2.2e-16

II.II. F-test

The F-test between the categorical feature LICNSTAT and response variable TOTALPMT_ADJ also results in effectively zero p-value, thus we can conclude significance between the two variables. TOTALPMT_ADJ is a continuous response variable of payment awarded for each case adjusted to April 2021 CPI levels.

R Code:
model1 <- lm(TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME, data = df)
model2 <- lm(TOTALPMT_ADJ ~ PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME, data = df)
anova(model2, model1)

Analysis of Variance Table
Model 1: TOTALPMT_ADJ ~ PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME
Model 2: TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME

Output:
Res.Df RSS Df Sum of Sq F Pr(>F)
1 223173 1.0846e+17
2 223116 1.0452e+17 57 3.9463e+15 147.8 < 2.2e-16 ***

II.III. Probability Variance

The bar plot below visualizes, for each licensing state, how many cases result in payment compared to the total number of allegations. The ratios, which can be inferred as probabilities, are listed on the right hand side of the graph for each state.

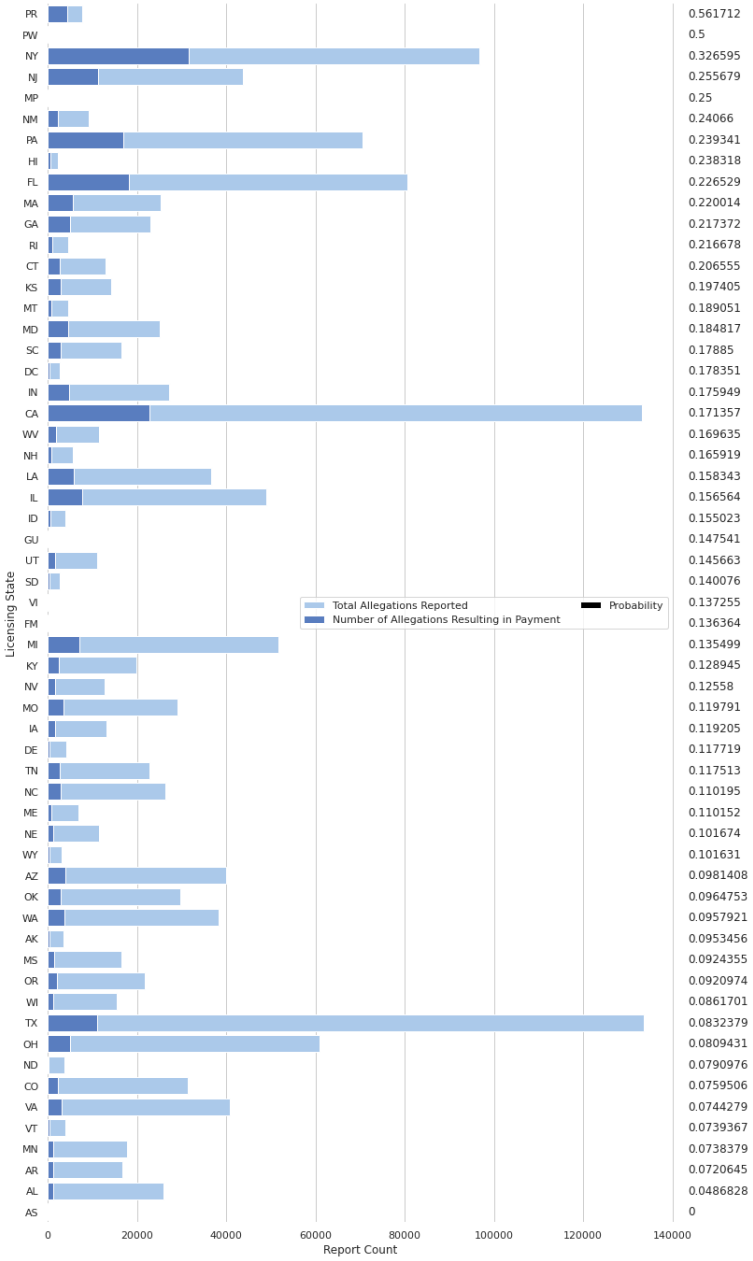


Figure 5: Probability of a malpractice payment accross licensing states

II.IV. Severity Variance

The box plot below visualizes how malpractice payment varies for each licensing state. The plot describes logged payment data adjusted to April 2021 CPI levels.

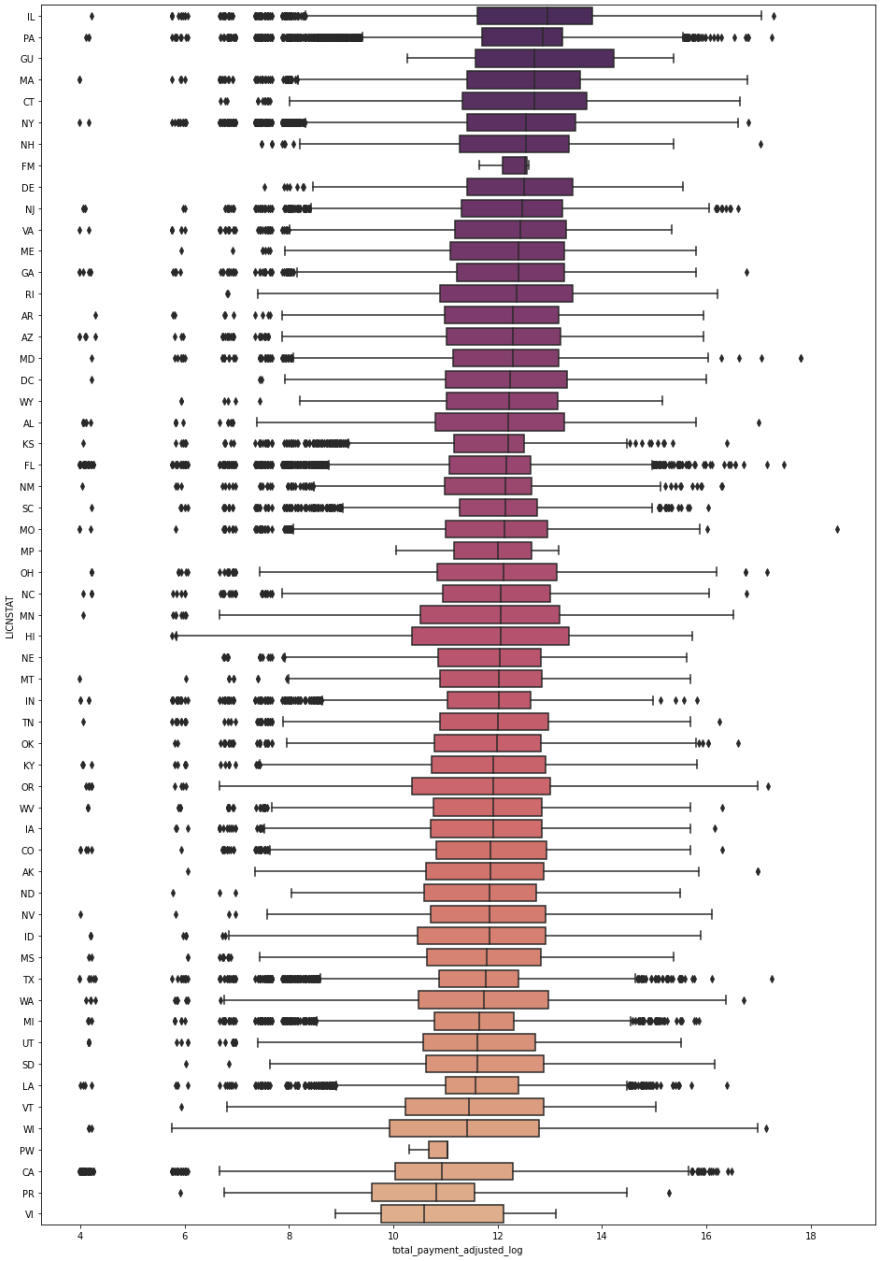


Figure 6: Log payment ranges accross licensing states

2.3.2 Feature 2: Practitioner's Age Group (PRACTAGE)

The selection of Practitioner Age as feature is based on correlation between practitioner age and experience, and thus is predictive of the probability and severity of a malpractice payment.

I. Feature Overview

The vast majority of the dataset is labeled with the practitioner's age. Among those, a significant majority falls between ages 21 and 50. An age group of 30 represents those with ages 21 to 30, 40 represent ages 31 to 40, and so on. Payment medians increase until age group 40 and decreases thereafter. It is highest at age group 40.

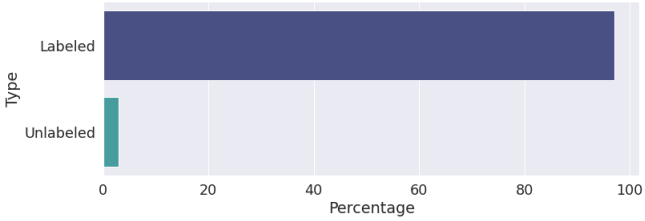


Figure 7: Almost all rows in the dataset are labeled with the associated practitioner's age

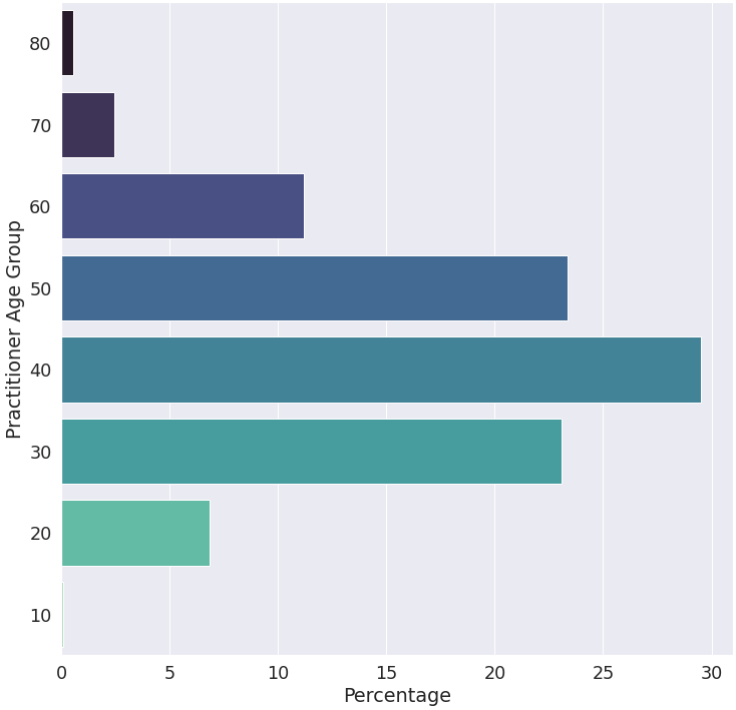


Figure 8: The majority of practitioners in dataset are between 21 and 50

II. Statistical Significance

As we did previously with Licensing State, we conduct Chi-sqared and F-test to estimate significance. We will also inspect visually the variance in probability and severity among practitioner age groups.

II.I. Chi-Square Test

The Chi-Square test between the categorical feature PRACTAGE and response variable PMT results in effectively zero p-value. This means we cannot reject the null hypothesis that there is no significance between the two variables.

R Code:
practage.data = table(df$PRACTAGE, df$PMT)
chisq.test(practage.data)

Output:
Pearson's Chi-squared test
data: practage.data
X-squared = 20057, df = 7, p-value < 2.2e-16

II.II. F-test

The F-test between the categorical feature LICNSTAT and response variable TOTALPMT_ADJ also results in effectively zero p-value, thus we can conclude significance between the two variables. TOTALPMT_ADJ is a continuous response variable of payment awarded for each case adjusted to April 2021 CPI levels.

R Code:
model1 <- lm(TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME, data = df)
model3 <- lm(TOTALPMT_ADJ ~ LICNSTAT + LICNFELD + ALEGATN1 + OUTCOME, data = df)
anova(model3, model1)

Analysis of Variance Table
Model 1: TOTALPMT_ADJ ~ LICNSTAT + LICNFELD + ALEGATN1 + OUTCOME
Model 2: TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME

Output:
Res.Df RSS Df Sum of Sq F Pr(>F)
1 223123 1.0453e+17
2 223116 1.0452e+17 7 1.0574e+13 3.2246 0.002023 **

II.III. Probability Variance

The bar plot below visualizes the ratio of cases which results in payment to the total number of cases among practitioner age groups.

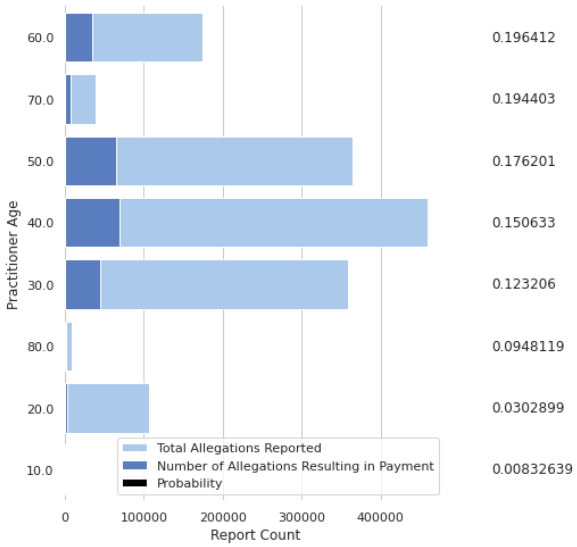


Figure 9: Probability of a malpractice payment among age groups

II.IV. Severity Variance

The box plot below visualizes how malpractice payment varies for each age group. The plot describes logged payment data adjusted to April 2021 CPI levels.

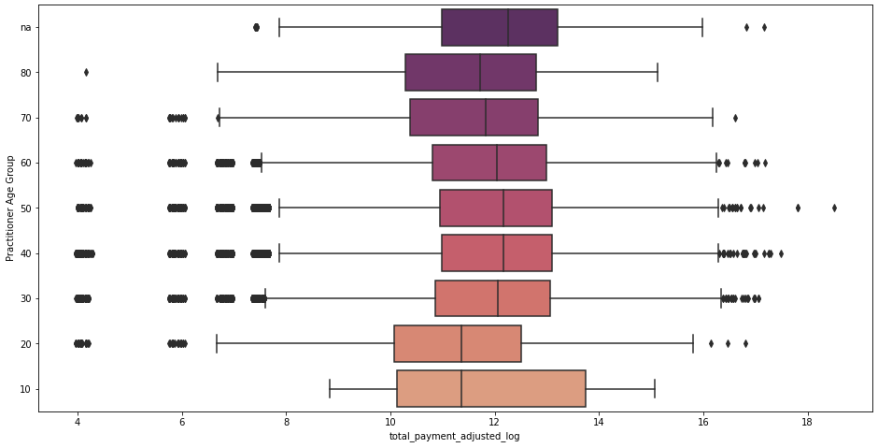


Figure 10: Log payment ranges among age groups

2.3.3 Feature 3: License Field (LICNFELD)

License field of the practitioner associated with a case most intuitively correlates with the level of risk of a procedure. For example, practitioners with either an Allopathic Physician (MD) or Osteopathic Physician (DO) license generally have authorizations to carry out higher risk procedures compared to Registered Nurses. The inherently higher risk translates to higher likelihood and severity of a malpractice payment for cases involving practitioners with those licenses.

I. Feature Overview

Virtually all cases in the dataset classify the license field of the practitioner involved. Cases involving those with licenses such as Allopathic Physician, Osteopathic Physician, Nurse Practitioner, and Physician Assistant have significantly higher likelihood of resulting in a payment compared to cases involving licenses such as Nurses, Therapists, or Counselors.

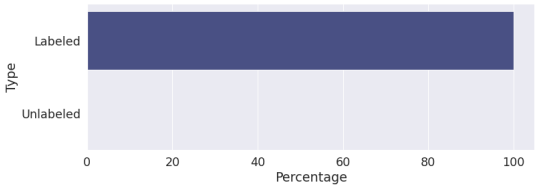


Figure 11: All cases are labeled with the license field of their associated practitioner

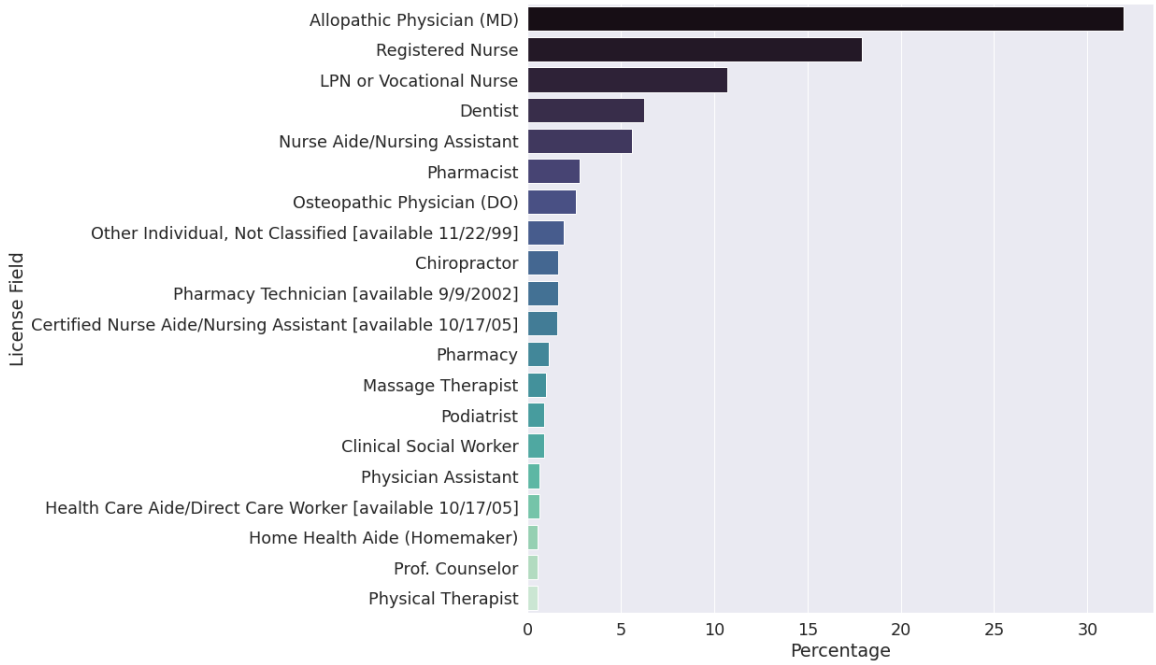


Figure 12: Top 20 most common license types in the dataset

II. Statistical Significance

II.I. Chi-Square Test

The Chi-Square test between the categorical feature LICNFELD and response variable PMT results in effectively zero p-value. This means we cannot reject the null hypothesis that there is no significance between the two variables.

R Code:
licnfeld.data = table(df$LICNFELD, df$PMT)
chisq.test(licnfeld.data)

Output:
Pearson's Chi-squared test
data: licnfeld.data
X-squared = 272834, df = 161, p-value < 2.2e-16

II.II. F-test

The F-test between the categorical feature LICNFELD and response variable TOTALPMT_ADJ also results in effectively zero p-value, thus we can conclude significance between the two variables.

R Code:
model1 <- lm(TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME, data = df)
model4 <- lm(TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + ALEGATN1 + OUTCOME, data = df)
anova(model4, model1)

Analysis of Variance Table
Model 1: TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + ALEGATN1 + OUTCOME
Model 2: TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME

Output:
Res.Df RSS Df Sum of Sq F Pr(>F)
1 223215 1.0501e+17
2 223116 1.0452e+17 99 4.9195e+14 10.608 < 2.2e-16 ***

II.III. Probability Variance

The bar plot below visualizes the ratio of cases which results in payment to the total number of cases among practitioner license fields.

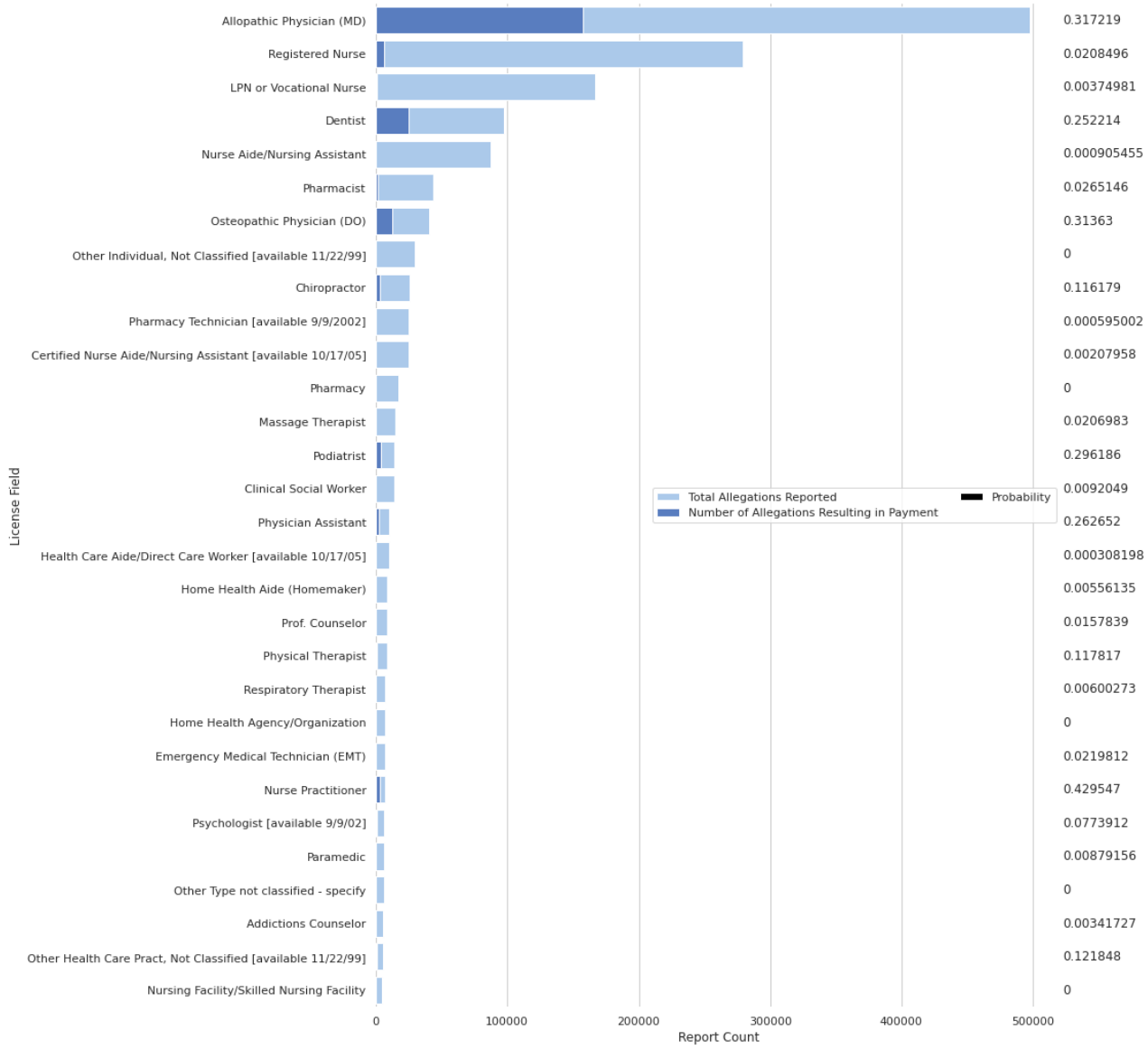


Figure 13: Probability of a malpractice payment among license fields

II.IV. Severity Variance

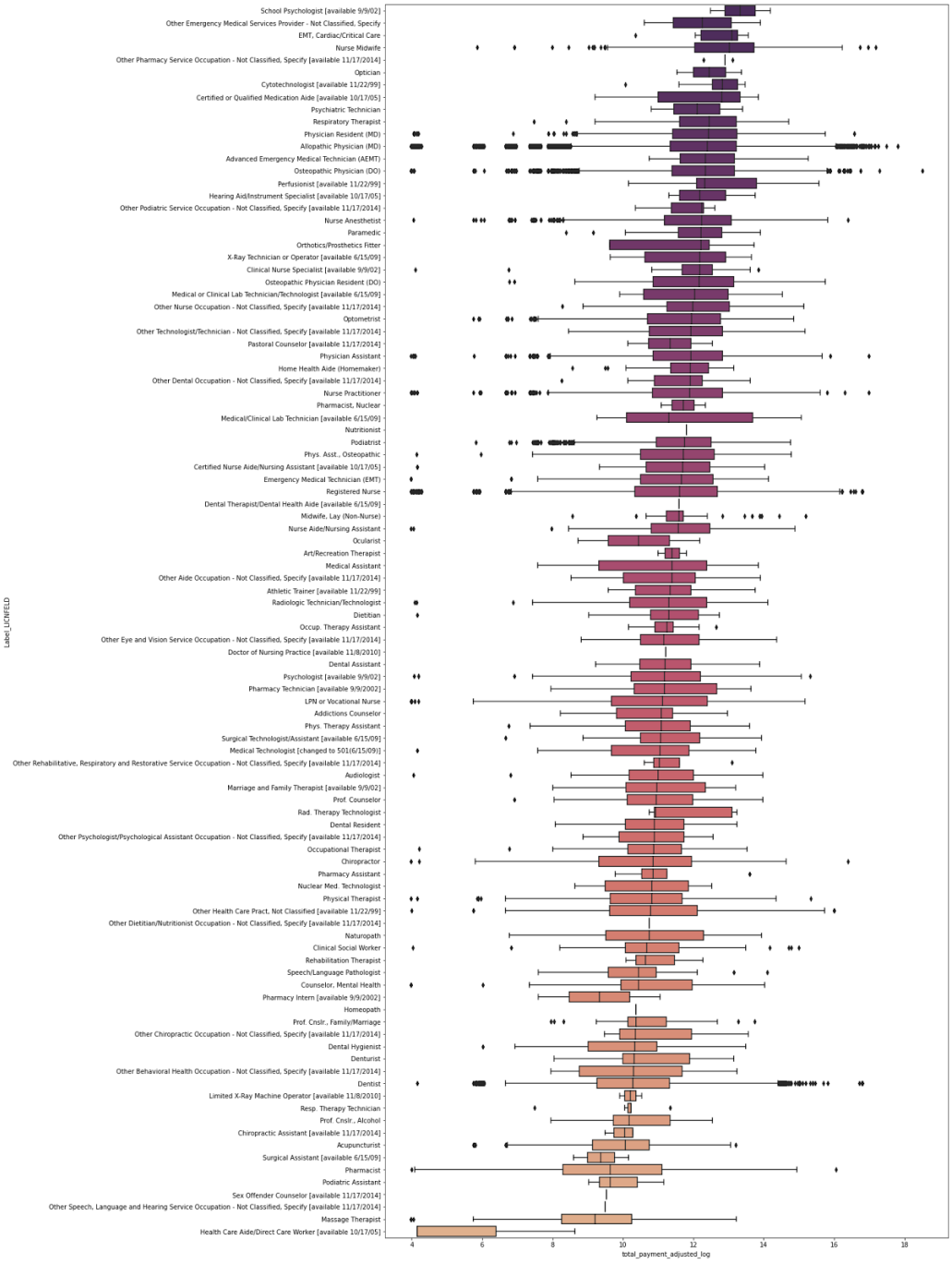


Figure 14: Log payment ranges among practitioner license fields

2.3.4 Feature 4: Primary Alegation Type (ALEGATN1)

Intuitively allegation types correlates most readily with the severity of a malpractice payment. "Failure to treat" or "delay in treatment" allegations for example, are intuitively more prone to be more severe than "unnecessary test".

I. Feature Overview

Only approximately 30% of the dataset is labeled with allegation type, and out of those the vast majority results in payment. Important note: there is a strong connection between those cases which results in payment and those with allegation type labels. This indicates that a case is far likelier to be given an ALEGATN1 label if it results in payment, which means the probability estimation given in probabilty variance visualization may not be indicative of the true probability.

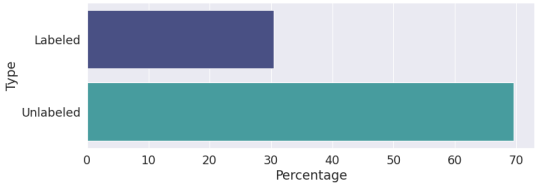


Figure 15: Approximately 30% of report includes information on alegation type

Failure to diagnose is the most common allegation type in the dataset, followed by improper performance and improper management.

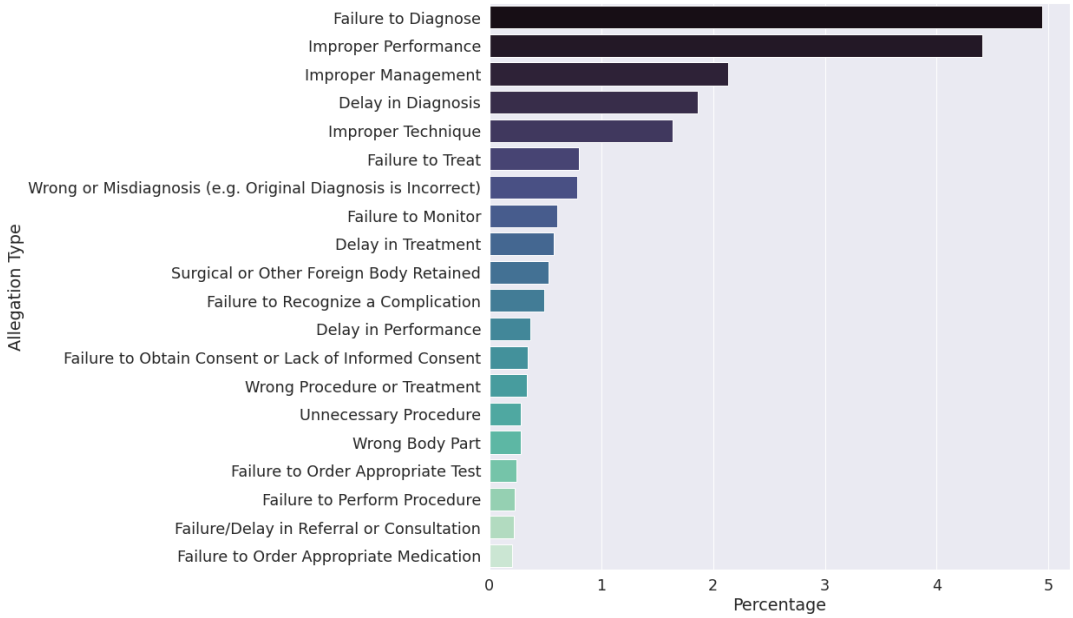


Figure 16: Top 20 most common allegation types in the dataset

II. Statistical Significance

II.I. Chi-Square Test

The Chi-Square test between the categorical feature ALEGATN1 and response variable PMT results in effectively zero p-value. This means we cannot reject the null hypothesis that there is no significance between the two variables.

R Code:
alegatn1.data = table(df$ALEGATN1, df$PMT)
chisq.test(alegatn1.data)

Output:
Pearson's Chi-squared test
data: alegatn1.data
X-squared = 124999, df = 90, p-value < 2.2e-16

II.II. F-test

The F-test between the categorical feature ALEGATN1 and response variable TOTALPMT_ADJ also results in effectively zero p-value, thus we can conclude significance between the two variables.

R Code:
model1 <- lm(TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME, data = df)
model5 <- lm(TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + OUTCOME, data = df)
anova(model5, model1)

Analysis of Variance Table
Model 1: TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + OUTCOME
Model 2: TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME

Output:
Res.Df RSS Df Sum of Sq F Pr(>F)
1 223206 1.0513e+17
2 223116 1.0452e+17 90 6.106e+14 14.483 < 2.2e-16 ***

II.III. Probability Variance

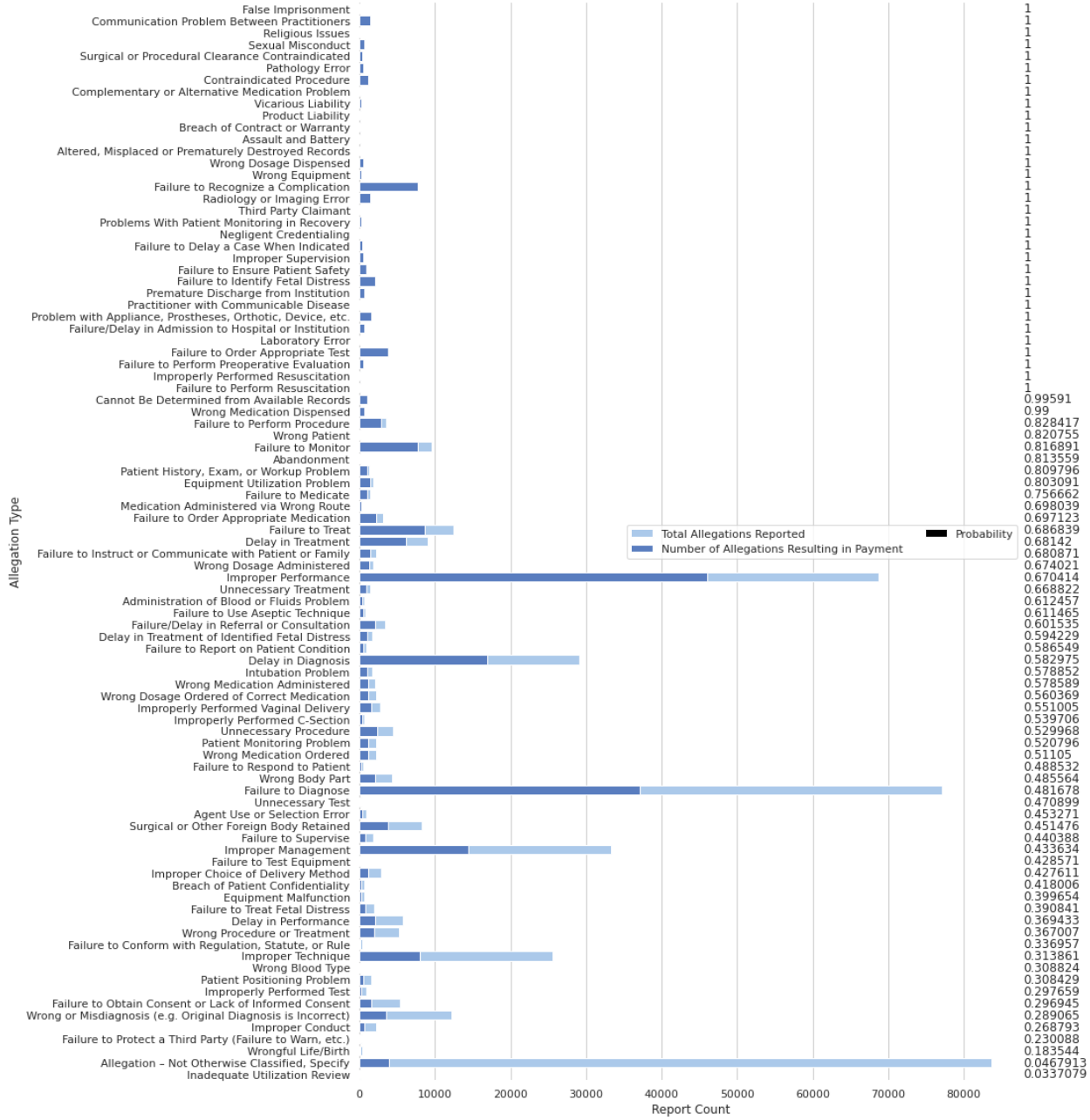


Figure 17: Probability of a malpractice payment among allegation types

II.IV. Severity Variance

Figure 18: Log payment ranges among practitioner allegation types

2.3.5 Feature 5: Case Outcome (OUTCOME)

The severity of the outcome of a procedure intuitively has a positive correlation with the severity of the resulting malpractice payment. .

I. Feature Overview

Less than 15% of all cases in the dataset are labeled with a specific outcome. Out of those labeled, all result in a payment to the alleging party. As a result, in the implementation of the model, OUTCOME will be used to calculate only severity and not probability (as there is no sample of a case with a known outcome that do not result in payment).

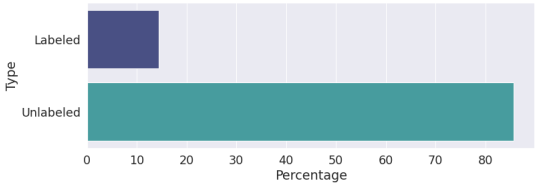


Figure 19: 15% of all cases in the dataset are labeled with a specific outcome. All cases with an OUTCOME label result in a payment.

The graph below indicates that the highest malpractice payments belong to those cases which result in permanent injuries, which may also require continuing care to the patients

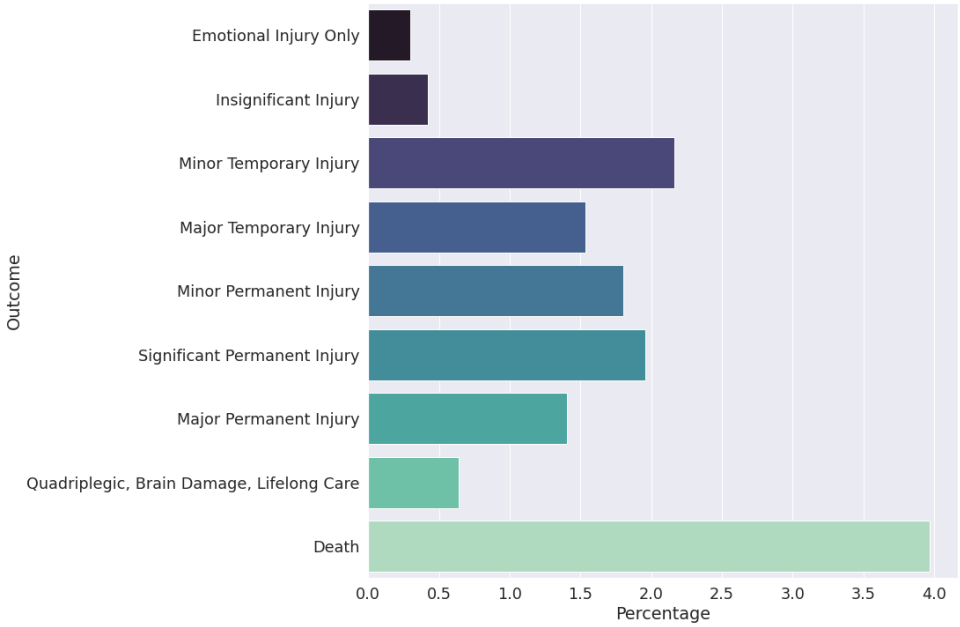


Figure 20: Distribution of outcomes in the dataset

II. Statistical Significance

II.I. Chi-Square Test

As we will not be using the OUTCOME feature for probability estimation, a Chi-squared test is unnecessary.

II.II. F-test

The F-test between the categorical feature OUTCOME and response variable TOTALPMT_ADJ also results in effectively zero p-value, thus we can conclude significance between the two variables.

R Code:
model1 <- lm(TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME, data = df)
model6 <- lm(TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1, data = df)
anova(model6, model1)

Analysis of Variance Table
Model 1: TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1
Model 2: TOTALPMT_ADJ ~ LICNSTAT + PRACTAGE + LICNFELD + ALEGATN1 + OUTCOME

Output:
Res.Df RSS Df Sum of Sq F Pr(>F)
1 223125 1.1314e+17
2 223116 1.0452e+17 9 8.6223e+15 2045.2 < 2.2e-16 ***

II.III. Severity Variance

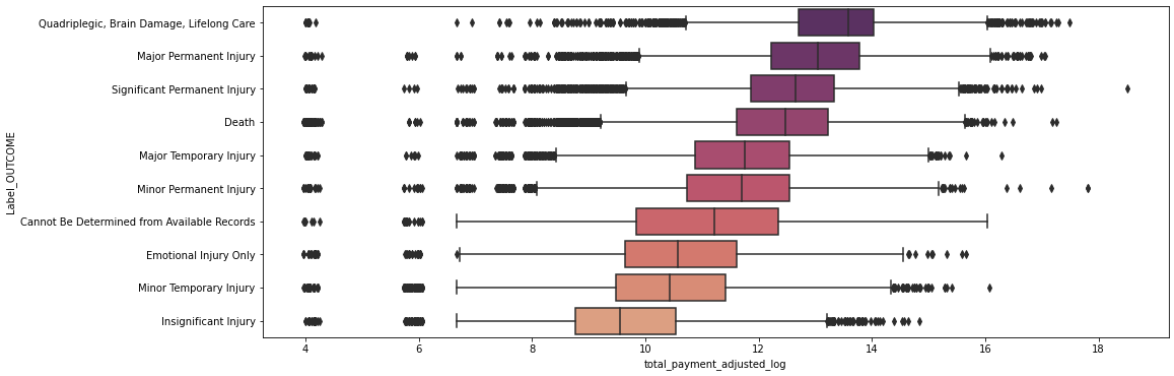


Figure 21: Log payment ranges among outcomes

3 Probability Component Methodology

3.1 Overview

The purpose of the probability model is to predict whether or not a claim will result in a payment. As such, they are binary classification models which, given the inputs of a practitioner's licensing state, age, license field, and the case's allegation type, yields either a positive or negative output. The aim of the model building process is to find a model with high accuracy, as measured on the hidden test set.

3.2 Methodology

3.2.1 Data Preprocessing

The binary target variable 'PMT' was derived from 'TOTALPMT' column of the original dataset, a simple labeling of '0' if the there is no payment associated with a case and '1' if the case results in a total payment of greater than zero. As all the features selected for the model are strictly categorical, one-hot encoding was used to translate the four features into over 300 features initially. Several license field features which are non-medical were removed, as well as the feature of state of American Samoa, which results in 255 one-hot encoded features used in building the model.

3.2.2 Train and Test Split

An 80-20 test split was used. The splitting process ensured that the train and test set retains the ratio of positive to negative cases in the original dataset. In other words, in both the training set (80%) and test set (20%), 14.38% are positive cases. This was done by first separating the positive and negative cases, and performing a random 80-20 split on each. The 80 percent split of both positive and negative cases were then joined to form the training set, and the remaining 20 percent of positive and negative cases formed the hidden test set.

3.2.3 Hyperparameter Tuning

Hyperparameter Tuning was performed on the Random Forest, Ada Boost, and Neural Net classifiers. To reduce search space, default hyperparameters such as 'ReLu' for activation function in Neural Net and 'adam' for solver in Random Forest were used. GridsearchCV was used to search for numeric hyperparameters such as n_estimators and max depth.

3.3 Results

The following are the resulting accuracy scores of each classifier evaluated against the hidden test set:

Naive Bayes Classifier: 88.26%
Random Forest Classifier: 91.46%
Neural Net (MLP): 91.84%
AdaBoost: 90.01%

3.4 Limitations

Imbalanced dataset limits the model significantly. As positive cases make up only 14% of the data, the models are less capable of predicting positive cases. Another significant limitation is the true probability estimation of a claim resulting in payment is unknown. In many cases, a probability estimation is a more valuable output than a classification.

4 Severity Component Methodology

4.1 Overview

The model's severity component seeks to estimate the amount of payment *given* that a lawsuit results in a payment. This means we include only cases which result in payment in our dataset to build the component. Recall that 223,926 rows out of 1,557,701 rows of the NPDB dataset resulted in a payment.

4.2 Methodology

In building this component we apply a process widely followed within the data science industry, starting from data preprocessing and concluding in model evaluation and selection. We apply standard procedures such as train-test split, hyperparameter tuning, and k-fold cross validation, with the objective of finding the model which best predicts the severity of a medical malpractice payment given the user's inputted variables. For simplicity, we use minimization of mean-squared error on the test set as the sole model selection criteria.

4.3 Data Preprocessing

Outlier Removal. We remove outliers to reduce the noise in our models. In this component, we do so by eliminating data with payment which falls outside the .05 and .95 quantiles. This translates to removing all rows with payments less than $8,800 or more than $1,250,000. The exclusion of this subset leaves still over 200,000 rows of data while reducing the average mean-squared error from $650,000 to $250,000 when applied to the random forest regressor fit, as will be explained in later section.

Train-test split. We apply the conventional 80-20 split to train and test the model. The 80% train set will be used also for hyperparameter tuning for each model. We use K-fold cross-validation (K = 5) on the training set for both hyperparameter tuning and loss-calculation. As we use K-fold cross-validation to perform hyperparameter tuning there is no need to split the data further for a validation set. The 20% test set will be used entirely to simulate unseen data, and will be used only during final evaluation and model selection.

4.4 Model Selection and Evaluation

4.4.1 Linear Regression

We begin our analysis with a linear regression model. Additionally, we fit a ridge and lasso model with alpha determined through using gridsearchCV.

I. Hyperparameter Tuning

We conduct hyperparameter tuning for both lasso and ridge in order to find the optimal alpha. We set the search space for alpha to equal 1, 5, 10, 20, and 50. The optimal alpha is 50 for ridge and 10 for lasso.

Python Code:

```
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge

rdg = Ridge()
lss = Lasso()

param = {'alpha':[1, 5, 10, 20, 50]}

GS_object = GridSearchCV( rdg, param, scoring='neg_mean_squared_error', cv = 5, )
GS_object = GridSearchCV( lss, param, scoring='neg_mean_squared_error', cv = 5, )
```

II. Root Mean-Squared-Error (RMSE) Estimation

We estimate RMSE using K-fold cross validation on 80% of our data found in train df. We set K = 5.

Python Code:

```
lm = LinearRegression()
n_splits = 5
cv = KFold(n_splits)


errors = []
for train_index, test_index in cv.split(X):
    X_train_kfold, y_train_kfold = X.iloc[train_index], y.iloc[train_index]
    X_test_kfold, y_test_kfold = X.iloc[test_index], y.iloc[test_index]
    lm.fit(X_train_kfold, y_train_kfold)
    y_pred = lm.predict(X_test_kfold)
    errors.append(np.sqrt(mean_squared_error(y_test_kfold, y_pred)))
```

The above evaluation yields:

errors:
256419.84385517793,
255454.911203295,
252929.77293623812,
253868.18861559324,
256805.52590108235

And the average RMSE is 255,095.65

We apply the same procedure to both ridge and lasso models with alpha values dervied from grid-searchCV to arrive at the following:

rdg_errors:
256166.83590415915,
255382.51760220042,
252870.58033655697,
253749.86784681273,
256717.39026046454

and average RMSE for the above ridge errors is 254,977.44

lss_errors:
256171.96440199565,
255392.55999019698,
252865.96220085118,
253750.28987366968,
256716.49697749183

and average RMSE for the above lasso errors is 254,979.45

III. Final Evaluation

We finally fit the optimal ridge and lasso models with the entire training data and evaluate it against the unseen 20% test data separated during data preprocessing step.
rdg_optim.fit(X, y)
y_pred_rdg = rdg_optim.predict(X_test)
np.sqrt(mean_squared_error(y_test, y_pred_rdg))

RMSE = 255086.4104747788

lss_optim.fit(X, y)
y_pred_lss = lss_optim.predict(X_test)
np.sqrt(mean_squared_error(y_test, y_pred_lss))

RMSE = 255092.88153757234

IV. Conclusion

The ridge and lasso errors suggest the models estimate claim payments with an average error of approximately 250,000 dollars. In the following sections, we follow the same procedures of tuning and cross-validation on several other models. For brevity, only the training errors and final test set errors are included. These models also indicate errors of approximately 250,000 dollars.

4.4.2 Random Forest Regressor

I. CV Training Erors
errors:
260736.34
253813.54

251964.04
252129.17
257519.53

Average RMSE = 255,232.53

II. Final Test Set Evaluation
RMSE = 253343.85143545613

4.4.3 MLPRegressor

I. CV Training Erors
errors:
261637.9309224144,
269475.79050049215,
262130.96203249734,
259725.62088146273,
262474.88955083914

Average RMSE = 263089.04

II. Final Test Set Evaluation
RMSE = 260767.14139376793

4.4.4 AdaBoost Regressor

I. CV Training Erors
errors:
328415.50777647435,
325285.58628037,
327781.0433801931,
325094.5360263175,
326751.6349728988

Average RMSE = 326,665.66

II. Final Test Set Evaluation
RMSE = 330424.50240746647

4.5 Limitations

The primary limitation of the severity models is the lack of stochastic estimations of the payment distributions. We can further develop the models by assuming a certain distribution (i.e. normal distribution) and conduct tests to see whether assumptions such as normality, linearity, and homoscedasticity are valid. A loss distribution then can be generated which allows users of the model estimate the estimated payment ranges of a particular case.