**UNIVERSITY OF ILLINOIS AT CHICAGO**

**IDS 575 – MACHINE LEARNING & ADVANCED STATISTICS**

**FINAL REPORT**

**HEALTHCARE FRAUD DETECTION ANALYSIS**

Group – 16

*NAME UIN*

*Changlin Yang 661053025*

*Preethi Srinivasan 663981973*

*Nissy Joy 659382267*

*Yang Zhang 650084008*

ABSTRACT

Healthcare fraud is a very expensive, crime in the United States, and it is not a victimless crime. Costs associated with the frauds are passed on to the population in the form of increased premiums or serious harm to beneficiaries. The complexity of the data systems and varied health models across the US, implementing digital advancements in healthcare is difficult. Fraud management is divided into two goals: fraud prevention and fraud detection. Fraud prevention in healthcare is defined as any action or policy that is in place to prevent any system abuse. Fraud detection, on other hand, is defined as identifying fraud as early as possible once a fraudulent scheme has already been perpetrated. The end goal of healthcare fraud detection is to provide leads to the investigators that can then be inspected more closely with the possibility of recoveries, or referrals to the appropriate authorities or agencies. Here using machine learning algorithms, we have performed predictive modeling on our dataset. We used different ML models like Naïve bayes, SVM with different kernels, Logistic regression etc. and found Linear SVM to be the best model that would be suitable for our Healthcare dataset which predicts Fraud and Non-Fraud. We chose Linear SVM based on F1-score and AUC scores.

INTRODUCTION

**1.1 MOTIVATION**

Each year trillions are spent on Healthcare in the United States, representing billions in insurance claims. Although a small fraction, the fraudulent claims carry a high price tag financially and how they impact the integrity of our healthcare system. Caring for health has become more expensive, making both private and public administrators more cost conscious in recent years. Therefore, health decision-makers are actively looking for ways to reduce costs. One such way of saving potentially billions of dollars is to avoid and detect healthcare fraud. Detecting potential fraudulent providers or physicians can help save costs for patients as insurance can increase the premium to be paid.

**1.2** **PROBLEM STATEMENT**

Here we took the dataset from Kaggle. The dataset is not a single dataset but consisted of 4 different datasets like inpatient data, Outpatient data, Beneficiary data, and providers data, combining all we formed the final combined dataset which we used for our predictive modeling. Healthcare fraud and abuse take many forms. Some of the most common types of frauds by providers are:

a) Billing for services that were not provided.

b) Duplicate submission of a claim for the same service.

c) Misrepresenting the service provided.

d) Charging for a more complex or expensive service than was provided.

e) Billing for a covered service when the service provided was not covered.

The goal of this project is to " predict the potentially fraudulent providers " based on the claims filed by them, along with this, we will also discover important variables helpful in detecting the behavior of potentially fraud providers. further, we will study fraudulent patterns in the provider's claims to understand the future behavior of providers.

RELATED WORK

* 1. **BACKGROUND AND RELEVANT WORK**

Some of the previous works in this field include an on-line discounting learning algorithm to identify whether a case has a high probability of being a statistical outlier and was used in identifying meaningful rare cases in health insurance pathology data from Australia’s Health Insurance Commission (HIC). Some studies have built a k-Nearest Neighbor algorithm with distance metric being optimized, a multi-layer perceptron (MLP) to classify practice profiles into four classes ranging from abnormal to normal.

The National Health Care Anti-Fraud Association[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9013219/#B1) conservatively estimates that about 3 percent of our healthcare spending is lost to fraud ($300 billion approximately) yearly. Fraud is a complex and difficult problem. It is important to acknowledge that fraud schemes constantly evolve, and fraudsters adapt their methods accordingly. The earliest account[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9013219/#B2) of “fraud” in the healthcare literature is from the 1860s when railway collisions were a frequent occurrence, leading to a controversial condition called “railway spine,” which later became a leading cause of personal injury compensation in rail accidents.

Healthcare fraud has evolved in the 21st century and has a varied set of profiles ranging from simple fraud schemes to complex networks. There is a vast amount of literature[8](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9013219/#B8), [9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9013219/#B9) available on fraud management techniques and models in different industries, such as healthcare, telecommunications, credit card services, insurance, and finance. Fraud prevention in healthcare can be defined as any action or policy that is in place to prevent any system abuse. For example, there is a Medicaid policy in the state of Texas[11](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9013219/#B11) for outpatient mental health services where certain types of providers, such as psychologists and licensed professional counselors, are limited to billing a combined maximum of 12 hours per day, regardless of the number of patients seen. This policy requirement is in effect to prevent fraud before it occurs.

Fraud detection is defined as identifying fraud as quickly as possible once a fraudulent scheme has already been perpetrated. Several articles discussed healthcare fraud data-mining methods in the literature with similar goals but from different perspectives, categorized the three different actors in healthcare fraud namely, providers, patients, and the payers and focused on the provider fraud literature. They further highlighted the scarcity in the data pre-processing methods from raw claims datasets to flattened datasets and commented on the importance of this step in identifying healthcare fraud using supervised and unsupervised methods. They also highlighted the two main types of classifier performance metric categories; 1) the error-based methods and 2) the cost-based methods, with error-based classifiers being more common in healthcare fraud literature.

The most recent review by Ai et al. (2021) discussed medical fraud detection methods in the literature using qualitative methods. They provided a methodological literature search using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines on the methods, number of peer-reviewed articles and a qualitative analysis of statistical methods, model performance, using evaluation metrics (when available) for health care domain. Their research is quite comprehensive, with a focus on being able to assess the strength of model performance and accuracy of existent fraud detection methods in the literature. They concluded that the evidence to provide a consolidated best method to identify healthcare fraud was inadequate considering the literature models were applicable to different domains within healthcare and therefore not directly comparable. They also highlighted that there was no literature available to estimate the cost of investigations to estimate potential cost savings using a fraud detection model. Here based on our dataset, we have used metrics like accuracy, ROC, AUC scores to determine which model works best and suitable w.r.t performance evaluation.

EXPERIMENTAL RESULTS

* 1. **DATA DESCRIPTION:**

Our dataset consists of Inpatient claims, Outpatient claims and Beneficiary details data of each provider with provider codes. The basic statistics of the datasets are shown below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | # of features | # of examples (train) | # of examples (test) |
| Beneficiary Data | 25 | 138556 | 63968 |
| Inpatient Data | 30 | 40474 | 9551 |
| Outpatient Data | 27 | 517737 | 125841 |

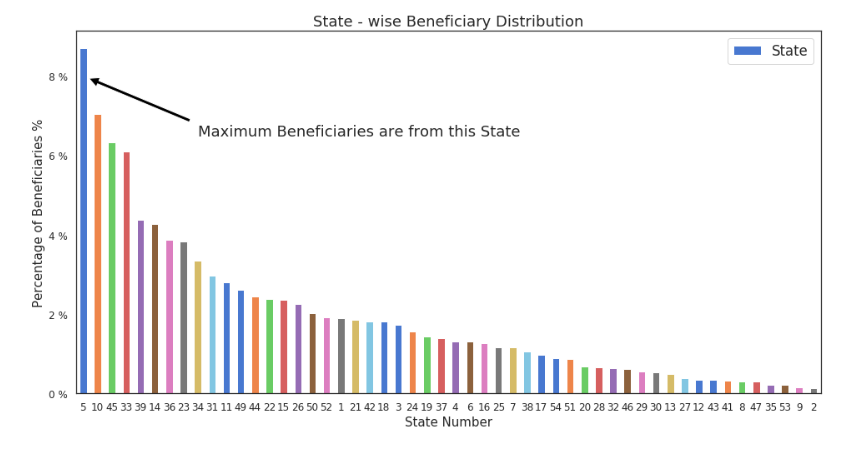
First, the Inpatient data provides insights about the claims filed for patients admitted in the hospitals, details like their admission, discharge dates and admit diagnosis code. Then, the Outpatient claims data provides details about the claims filed for patients who visit hospitals and are not admitted. And the Beneficiary data contains details like health conditions, region, state, race etc. In addition, the providers data consists of provider codes and if they are potentially fraudulent or not.

Some of the features of the datasets are shown below.

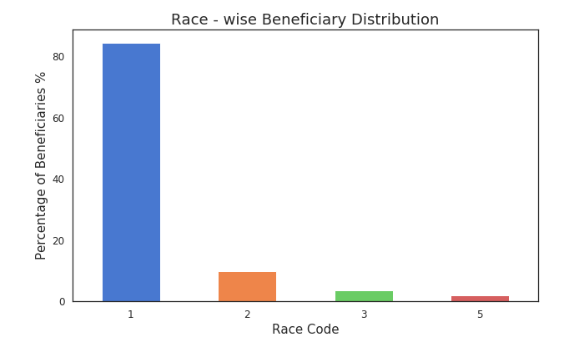
|  |  |
| --- | --- |
| Feature Name | Feature Description |
| IP & OP Data |  |
| Bene ID | The unique id of each beneficiary. |
| Claim ID | The unique id of the claim submitted by the provider. |
| AdmissionDt | The date on which the patient was admitted in the hospital in yyyy-mm-dd format. |
| DischargeDT | The date on which the patient was discharged from the hospital in yyyy-mm-dd format. |
| ClmProcedureCode \*  1,2,3,4,5,6 | The codes of the medical treatments or procedures performed by the provider or physicians for medication of a patient for a specific claim. |
| ... |  |
| Beneficiary Data |  |
| Bene ID | The unique id of the beneficiary. |
| DOB | The date of birth of the beneficiary. |
| DOD | The date of death of the beneficiary. |
| Gender | The gender of the beneficiary. |
| Race | The human race of the beneficiary. |
| Chronic Condition Indicators \*  1,2,3,4,5,6,7,8,9,10,11 | A code which indicates whether a beneficiary had a chronic condition of a specific disease at the time of buying a plan. |
| ... |  |
| Providers data |  |
| Provider | The unique id of each provider. |
| PotentialFraud | If the providers are potentially fraudulent or not. |

**1.2 EXPLORATORY DATA ANALYSIS:**

In the dataset collected from Kaggle, we have 82 features after combining all the three datasets viz beneficiary data, inpatient data, and outpatient data. After data pre-processing, we plotted the graphs between different features of the dataset to figure which features are important and how they will add value to our analysis. This will also help to discover patterns in the data, spot anomalies and check assumptions with the help of summary statistics and graphical representations.

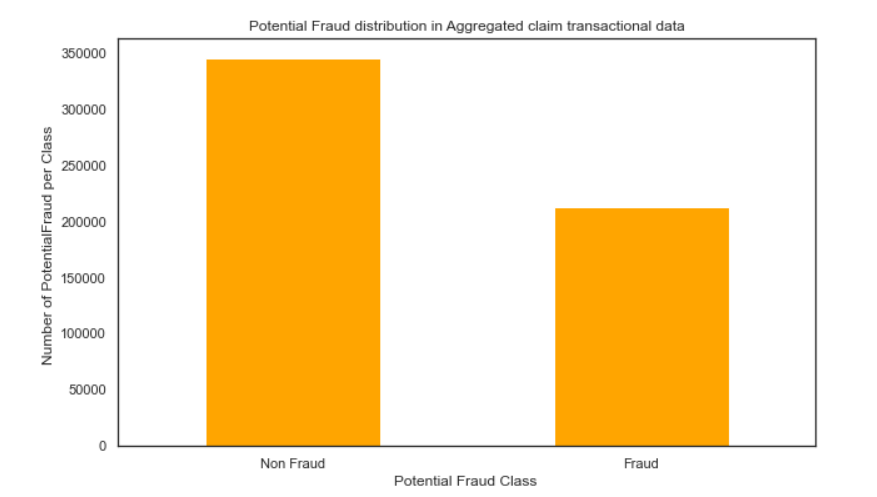


As we can see states 5, 10, 45 have the maximum percentage of beneficiaries when we plotted a graph of states vs percentage of beneficiaries.

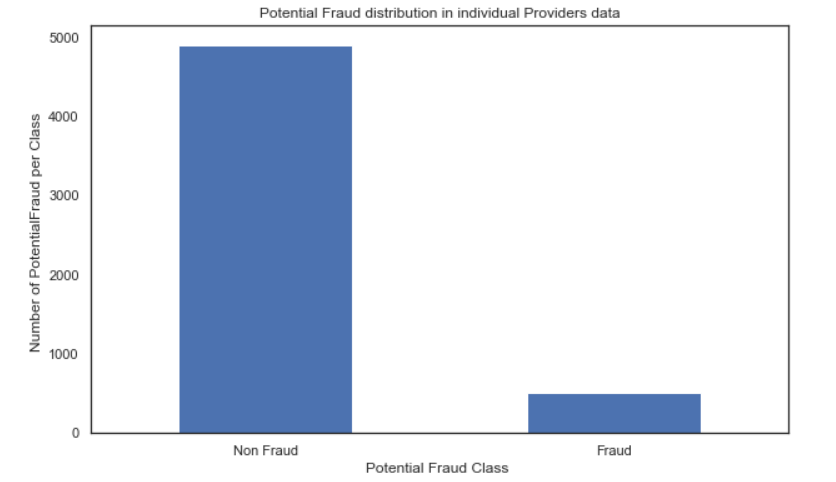


Race code 1 has the maximum percentage of beneficiaries as compared to other race codes 2, 3 5.

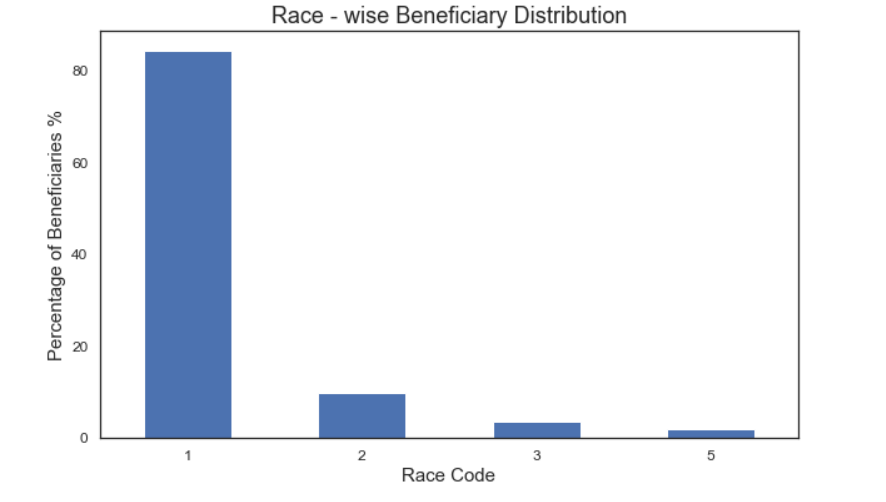
Here we plotted graph of the frequencies containing fraud and non-fraud Merged transactions in the data



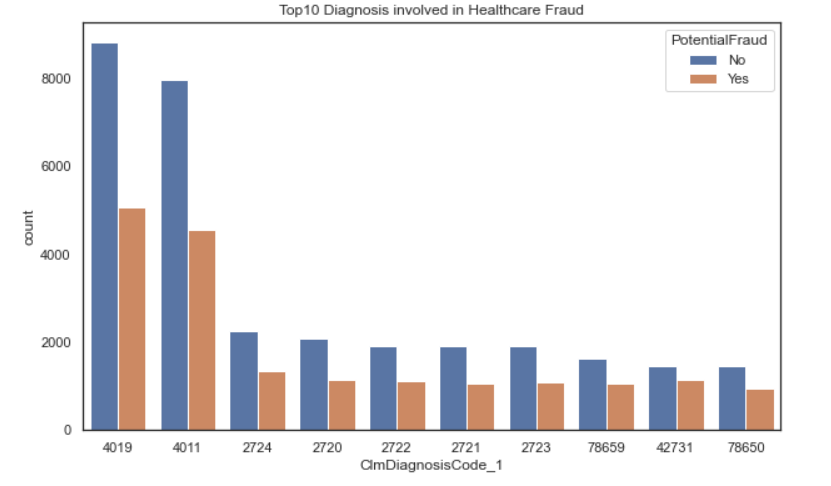
Here from the graph, we come to know that 38%, Percent Distribution of Potential Fraud class based on aggregated claims transaction data represents fraud class while 62% represents Non-Fraud class.



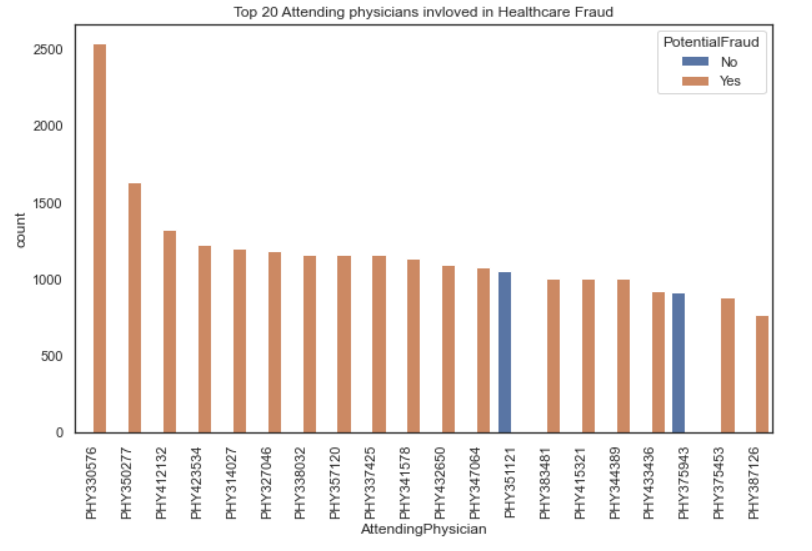
Here from the graph, we come to know that 9%, Percent Distribution of Potential Fraud class based on individual’s provider data represents fraud class while 91% represents Non-Fraud class. From the above 2 graphs, we can say that the proportion of fraudulent claim transactions are more compared to nonfraud providers. So, we must get insights from number of claim transactions and amounts involved per – Beneficiary, Beneficiary + Physician, Physician, Diagnosis, Procedure etc.



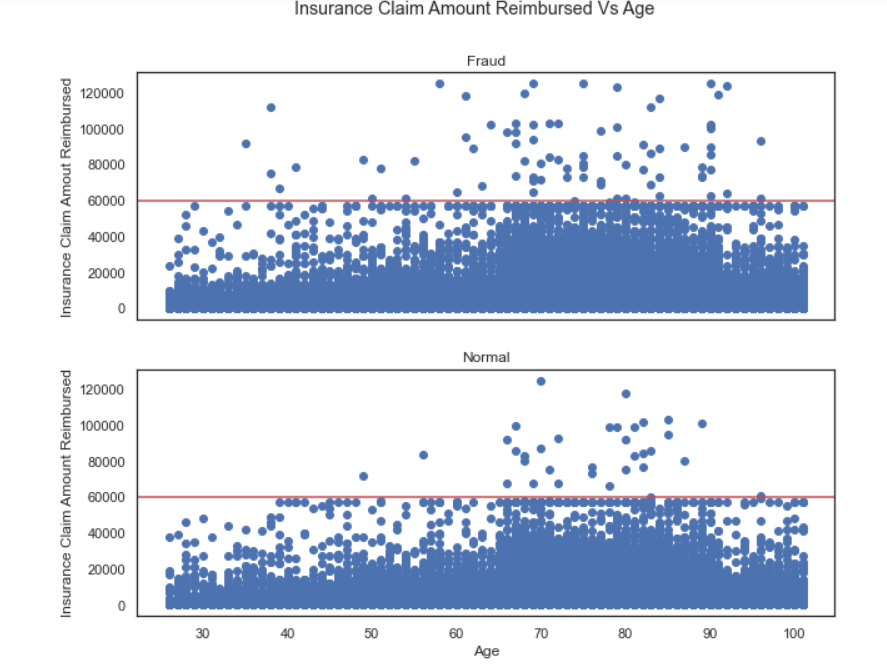
Here we have plotted the graph of frequencies of race-wise beneficiaries, with Race code 1 having the maximum percentage of beneficiaries.



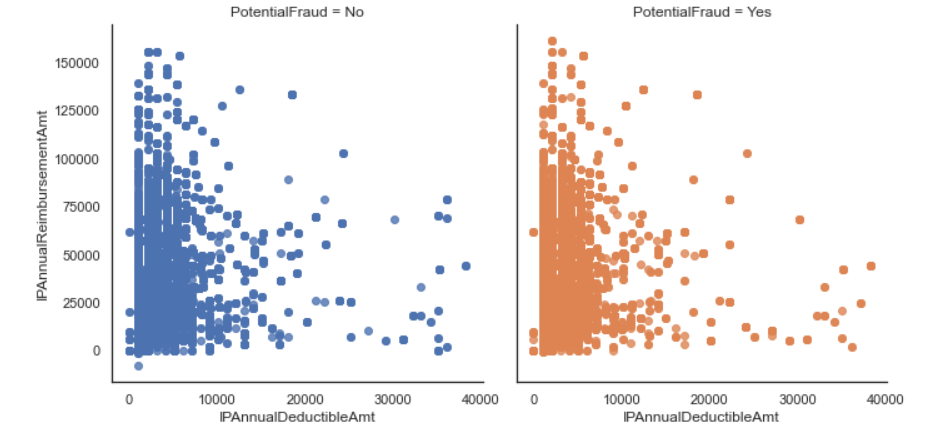
Here we have shown the plot of Top-10 Claim Diagnosis involved in Healthcare Fraud.



Here we have shown the plot of Top-20 Attending Physicians involved in Healthcare Fraud.



The above graphs show the Insurance Claim Amount Reimbursed Vs Age. From the above graph, we see that occurrence of fraud cases is more frequent in lower age groups (30-70 years) compared to higher age groups (70+ years). Age is one of the important features for differentiating between fraud and non-fraud behavior.



In the above graph, when we plotted a graph of IPAnnualDeductibleAmt vs IPAnnualReimbursementAmt, with potential fraud yes and no, we can see that, there is not difference that can be found between the two graphs, because of which we cannot come to a conclusion as to which feature would be more important for us w.r.t our predictive modeling and hence we will have to perform feature engineering to derive accurate and meaningful features.

**1.3 FEATURE ENGINEERING:**

Feature engineering is a machine learning technique that leverages data to create new variables that aren’t in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of**simplifying and speeding up data transformations**while also**enhancing model accuracy.**

Creating features involves creating new variables which will be most helpful for our model. This can be adding or removing some features. As we saw above, the cost per sq. ft column was a feature creation.

Feature transformation is simply a function that transforms features from one representation to another. The goal here is to plot and visualize data, if something is not adding up with the new features, we can reduce the number of features used, speed up training, or increase the accuracy of a certain model.

Feature extraction is the process of extracting features from a data set to identify useful information.

Here we performed Feature engineering, where we grouped the features, created some new features like “Age”, “Dead”, “Admitdays” and after performing feature engineering the train data has 188 features while the test data has 187 features. Further again we removed unnecessary columns like “ClaimID”,”ClaimStartDt”,”ClaimEndDt”,”AttendingPhysician” etc.

Other than basic explorations and visualizations, we can use certain methods to identify clues of fraud and abuse. One such simple method is 'Grouping based on Similarity'. In this method, we basically group all the records by the ProcedureCodes, DiagnosisCodes,Provider. We also performed feature engineering using combined features where we got more specific and accurate features.

After performing feature engineering, we did data pre-processing again. Performed Feature selection, removed unnecessary columns that were used to derive the new features like “BeneID”, “ClaimID”, “ClaimStartDt”, “ClaimEndDt” etc. Further checked for missing values and then aggregated the data to providers level in order to focus only on providers level data as we have to predict which providers are fraud and which providers are non-fraud.

Further performed train test split and standardization of the data to perform predictive modeling.

**1.4 MODEL DESIGN:**

Our goal is to build a binary classification model based on the claims filled by the provider along with inpatient, outpatient data and beneficiary data to predict whether the provider is potentially fraudulent or not.

For the classification problem we first started off with **Gaussian Naïve bayes model**. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of conditional independence between every pair of features given the value of the class variable. Bayes’ theorem states the following relationship, given class variable and dependent feature vector through :

Using the naive conditional independence assumption that

for all , this relationship is simplified to

Since is constant given the input, we can use the following classification rule:

and we can use Maximum A Posteriori (MAP) estimation to estimate and ; the former is then the relative frequency of class in the training set.

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of . “GaussianNB” implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

The parameters and are estimated using maximum likelihood.

Graphical user interface, text, application, email

Description automatically generated

Implementation of GaussianNB model for our data with prior probabilities for the classes

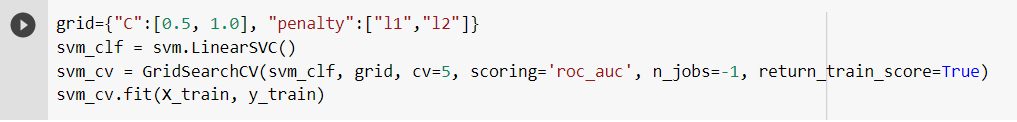
**Support Vector Classifier**:

Support vector machines (SVMs) are a set of supervised learning methods where we try to find a hyperplane that best separates the two classes. And can also be used for regression, and outliers’ detection. SVMs are mainly divided into two groups Linear SVM and Non-linear SVMs. When the data is perfectly linearly separable only then we can use Linear SVM. When the data is not linearly separable then we can use Non-Linear SVM, where some advanced techniques like kernel tricks to classify them. In most real-world applications we do not find linearly separable datapoints hence we use kernel trick to solve them.

The different kernels that are in SVMs can be specified though the kernel argument and their mathematical formulation is as follows:

* Liner:
* Polynomial: where is specified by parameter degree, by coef0
* rbf: where is specified by the parameter gamma, must be greater than 0
* sigmoid where is specified by coef0

For our classification problem we build one linear SVM model, and hyper tuned the parameters using the GridSearchCV algorithm.



Text

Description automatically generated

And trained our model with the best parameters identified which are “C” =1.0 and penalty=” l2”.

For non-linear SVMs we also applied the same GridSearchCV algorithm to identify the best parameters such as “C”, “gamma” and “kernel”.

Text

Description automatically generated with medium confidence

Text

Description automatically generated

We trained our SVM with the best parameters “C” = 1.0, “gamma” = 0.01 and “kernel” = “rbf”.

**Logistic Regression**:

Logistic Regression is a machine learning algorithm that allows us to create a classification model. The algorithm analyses independent variables and one dependent variable, to predict the output. This algorithm is used to predict categorical variables using independent variables which are continuous.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

The equation after applying sigmoid function to linear regression to achieve logistic regression:



We implemented logistic regression with libraries from sklearn as follows

Text

Description automatically generated

**1.5 MODEL EVALUATION:**

As our data is not balanced accuracy is not a proper measure of model evaluation, so we calculated AUC and F1-scores, confusion matrices at the best model threshold from AUC curves.

***Gaussian Naïve bayes Model:***

The gaussian naïve bayes model achieved an AUC of 0.7726 on validation data and an F1 score of 0.5669 at the best model threshold of 0.0469.

Chart, line chart

Description automatically generated

Graphical user interface

Description automatically generated

Naïve bayes model has the lowest AUC score out of all the models built, considerably as Naïve bayes assumes data follows a gaussian distribution with no-variance between dimensions but our data might have depended on variables leading to low evaluation metrics and misclassifications.

***Linear Support Vector Machine:***

Our linear-SVM model achieved an AUC of 0.9049 on validation data and an F1 score of 0.5301 at the best model threshold of 0.0699.

Chart

Description automatically generated

Chart, treemap chart

Description automatically generated

The linear-SVM has better AUC score than our baseline model of Naïve bayes model, and this model also has better balance between sensitivity (true positive rate) and specificity (true negative rate) than Naïve bayes.

***Non-Linear Support Vector Machine:***

Our non-linear SVM with the best hyper tuned parameters has achieved an AUC of 0.8075 on validation data and an F1 score of 0.5580 at the best model threshold of 0.0532.

Chart, line chart

Description automatically generated

Graphical user interface, treemap chart

Description automatically generated with medium confidence

The model AUC of SVM with “rbf” has dropped to 0.8075 from the linear-SVM model, we can deduce that our data is better classified linearly than non-linear.

***Logistic Regression:***

Our Logistic Regression model has achieved an AUC of 0.9358 and on validation data an F1 score of 0.5714 at the best model threshold of 0.3967.

Chart, line chart

Description automatically generated

Chart, treemap chart

Description automatically generated

The logistic regression the best AUC of all closely followed by our linear-SVM model and the specificity and sensitivity are also very similar.

Although we see a significant difference in the AUC evaluation metric from one model to another, if closely observed we hardly see any major difference in their F1-score. All the model F1-scores are low indicating poor model performance and data quality.

CONCLUSION

Our project is to build a machine learning model to detect possible fraud in health care and prevent it in time. We built many models and compared their performance, which are the Naive Bayes Model, Linear SVM, SVM, And Logistic Regression. Logistic Regression has the best performance concerning accuracy as a measure of 91.25%.

Although we have obtained a high accuracy rate, the model still has many shortcomings. Our highest F1 score was only 0.57, indicating that the performance of our model was not good enough to classify the results. Our model had a high specificity but a very low sensitivity deficit, which means that although our model could correctly label groups to be used for fraud, it was not accurate for those who would not be used for fraud. This has alleviated the problem of health insurance fraud to some extent, but it still needs to be improved. The reason for this may be that our data cleaning is not complete, for example, highly correlated independent variables have not been cleaned, thus affecting the prediction of results. We will try to modify this in the future, and we will also try new models such as decision trees to see if there are better results.

And there is huge scope for research to identify the best-suited metrics and better feature engineering to build more robust models with more data on the providers. And deeper research into the payoff of misclassification can also aid us in evaluating models with cost matrices and getting better results as a result.

References:

[1] Kumaraswamy N, Markey MK, Ekin T, Barner JC, Rascati K. Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead. Perspect Health Inf Manag. 2022 Jan 1;19(1):1i. PMID: 35440932; PMCID: PMC9013219.

[2] Waghade, Shivani S., and Aarti M. Karandikar. "A comprehensive study of healthcare fraud detection based on machine learning." International Journal of Applied Engineering Research 13, no. 6 (2018): 4175-4178.

[3] National Health Care Anti-Fraud Association. "HealthCare Fraud–A Serious and Costly Reality For All Americans." April2005 (2007).

[4] Waghade, Shivani S., and Aarti M. Karandikar. "A comprehensive study of healthcare fraud detection based on machine learning." International Journal of Applied Engineering Research 13, no. 6 (2018): 4175-4178.