**Using Machine Learning to Detect Medicare Fraud**

Ryan Mulcahey

Fairfield University

DATA6999 – Capstone: Business Analytics

Dr. Illya Mowerman

**Table of Contents**

[**Summary** 3](#_Toc141113533)

[**Introduction** 3](#_Toc141113534)

[**Literature Review** 4](#_Toc141113535)

[*What algorithms have been used to create models for detecting Medicare fraud?* 4](#_Toc141113536)

[*What predictors have been used to model detection of Medicare fraud?* 4](#_Toc141113537)

[**Results** 5](#_Toc141113538)

[**Methods** 6](#_Toc141113539)

[*Data Collection and Processing* 6](#_Toc141113540)

[*Data Dictionary* 6](#_Toc141113541)

[*Univariate Statistics – Numeric Features* 7](#_Toc141113542)

[*Univariate Statistics – Categorical Features* 8](#_Toc141113543)

[*Data Preparation* 8](#_Toc141113544)

[*Feature Selection* 9](#_Toc141113545)

[*Modeling* 9](#_Toc141113546)

[Aggregate Modeling Results 10](#_Toc141113547)

[**Citations** 10](#_Toc141113548)

[**Appendix** 10](#_Toc141113549)

[*Data Pipeline and EDA* 11](#_Toc141113550)

[*Feature Selector* 11](#_Toc141113551)

[*Links to Models* 11](#_Toc141113552)

[Naïve Bayes 11](#_Toc141113553)

[Logistic Regression 11](#_Toc141113554)

[Random Forest 11](#_Toc141113555)

[Decision Tree 11](#_Toc141113556)

[XGBoost 11](#_Toc141113557)

[XGBoost Grid Search 11](#_Toc141113558)

# **Summary**

Predicting Medicare fraud is becoming a more and more important issue, with changing demographics and resources becoming more constrained, fraud becomes a significant system resource drain preventing equitable care to the value it should be delivered at. Of the data available, several models were developed that focused on drivers of fraud at the provider level. With this, providers are grouped into risk categories for probability of committing fraud, where Medicare resources can be deployed to prioritize those in the highest risk category, allowing fraud investigators to allocate scarce manpower to achieve the highest rate of return on time invested.

# **Introduction**

The overarching objective of this research paper is to delve into the potential of machine learning in detecting and eventually preventing Medicare fraud. The task is an imperative segment of the broader agenda of modernizing and optimizing the healthcare industry through the power of artificial intelligence and data science. The paper aims to shed light on the problem of Medicare fraud, critically examine the role of machine learning in tackling the problem, and explore previous studies on this subject. Further, it intends to present a structured model integrating machine learning with a systematic detection mechanism that can revolutionize the future of Medicare fraud detection.

The significance of detecting and preventing Medicare fraud cannot be overstated. As of 2021, fraudulent activities in healthcare are estimated to cost the U.S. government approximately $68 billion annually, of which a substantial portion is attributed to Medicare. This not only financially drains a crucial public resource but also compromises the quality and integrity of the healthcare services provided to the citizens. The downstream effect of these fraudulent activities can lead to higher healthcare costs, less coverage for legitimate services, and a lack of trust in the system. Implementing a robust fraud detection system is imperative for the health of the healthcare sector and the welfare of the beneficiaries of these services.

The potential of machine learning in detecting Medicare fraud has been examined in previous studies. Simbert (2023) does a meta-analysis of 39 studies and concludes that machine learning is a powerful tool for fighting Medicare fraud and should be implemented in several ways. Alanzi (2022) explores various machine learning models and how they can be used to prevent or detect fraud. There is an abundance of literature available in 2023 that studies how machine learning and AI can help fight fraud, as the field is accelerating at a rate not seen before.

The structure of this paper seeks to approach the subject matter by presenting a clear layout of the relevant sections. The paper will begin with an introduction, providing a thorough understanding of Medicare fraud and its implications, followed by a detailed literature review analyzing previous studies on the subject. Subsequent to that, the paper will discuss the data to be used for the purpose of the study, detailing its relevance and validity. The subsequent sections will present the modeling approach, detailing the machine learning models and techniques utilized for fraud detection. The paper will conclude with a summary of findings, the limitations of machine learning in detecting Medicare fraud, and the implications of these findings for the future.

# **Literature Review**

## *What algorithms have been used to create models for detecting Medicare fraud?*

Medicare fraud is a significant issue that costs taxpayers billions of dollars and puts beneficiaries' health and welfare at risk. To address this problem, various machine learning algorithms have been applied to model and predict Medicare fraud.

One approach is to use supervised machine learning methods with three classification models, namely decision tree, random forest, and boosted tree, to detect Medicare fraud (Purnita & Mondal, 2021). They found that these algorithms were suitable for fraud detection in Medicare systems. Similarly, they stated that the random forest model has a competitive edge and is often superior in the prediction and classification compared to traditional logistic regression methods and other machine learning methods (Lin et al., 2022).

Another approach is to use unsupervised machine learning algorithms. used the Isolation Forest algorithm, an unsupervised machine learning algorithm, to detect Medicare fraud based on outliers (Kanksha et al., 2021). They found that this algorithm improved overall performance and achieved a high accuracy of 98.76% in detecting fraud.

In addition to traditional machine learning algorithms, deep learning methods have also been explored for Medicare fraud detection. compared six deep learning methods designed to address the class imbalance problem in Medicare fraud detection (Johnson & Khoshgoftaar, 2019). They used data-level techniques such as random over-sampling and random under-sampling, as well as algorithm-level techniques such as cost-sensitive loss functions. They achieved desirable class-wise performance by identifying optimal decision thresholds for each model.

Additionally, there have been studies on the use of statistical models and ensemble techniques for Medicare fraud detection. Vidmar et al. developed control limits for a double-square-root chart based on prediction intervals from regression-through-origin to detect DRG outliers (Lin et al., 2022). Bauder et al. invented a general outlier detection model based on Bayesian inference using probabilistic programming to detect claims fraud in Medicare medical insurance payments (Lin et al., 2022). A novel ensemble supervised feature selection technique to build explainable machine learning models for Medicare fraud detection (Hancock, 2023).

Overall, a variety of algorithms have been used to model and predict Medicare fraud, including supervised machine learning algorithms such as decision trees, random forests, and boosted trees, unsupervised machine learning algorithms such as the Isolation Forest algorithm, deep learning methods, statistical models, and ensemble techniques. These algorithms have shown promising results in detecting Medicare fraud and have the potential to improve fraud detection and reduce healthcare costs.

## *What predictors have been used to model detection of Medicare fraud?*

Medicare fraud detection models utilize various predictors to identify fraudulent activities within the Medicare system. Several studies have explored the selection of predictors in machine learning models for Medicare fraud detection.

One study by Johnson & Khoshgoftaar (2019) focused on the class imbalance problem in Medicare fraud detection. They compared six deep learning methods and evaluated data-level techniques such as random over-sampling (ROS), random under-sampling (RUS), and a hybrid ROS-RUS. The algorithm-level techniques evaluated included a cost-sensitive loss function, the Focal Loss, and the Mean False Error Loss. While this study did not explicitly mention the predictors used, it highlighted the challenges associated with class-imbalanced big data and the need to improve existing results.

Another study by Perols (2011) compared the performance of different statistical and machine learning models in detecting financial statement fraud. Out of 42 predictors examined, six predictors were consistently selected and used by different classification algorithms: auditor turnover, total discretionary accruals, Big 4 auditor, accounts receivable, meeting or beating analyst forecasts, and unexpected employee productivity. These findings can be extended to Medicare fraud detection models as they provide insights into predictors that can improve fraud risk models.

In a study by Bauder et al. (K & Ilango, 2019), the authors proposed a novel method for detecting fraud in health insurance, including Medicare. They used multiple predictors as model input to detect outliers in payment data. While the specific predictors were not mentioned, the study highlighted the effects of the class imbalance problem on machine learning performance.

Furthermore, Lin et al. (2022) mentioned that CatBoost, a machine learning algorithm, is commonly utilized in Medicare fraud detection. However, the specific predictors used in their study were not mentioned.

More research is needed to explore and identify additional predictors that can improve the performance of Medicare fraud detection models.

# **Results**

Given a dataset from 2008-2009 calendar years, several models were run against this data, with a Random Forest model providing what seems like the best approach. Here we found some major drivers that would predict Medicare fraud in descending importance: maximum reimbursement for outpatient visits, patient age, amount reimbursed for a specific claim, maximum reimbursement for inpatient visits, patient risk score and claim period in length.

Decision tree used a very similar predictor set of features, with different weights that ended up with a slightly lower performing output. Logistic regression, naïve bayes and XGBoost both keyed in on one major variable: hospitalization duration. While I think this is could be a strong driver, I believe other factors can influence occasions where fraud is committed.

Limitations include the data this research is formed upon. While the feature set includes up to 50 features after feature engineering, there may be a significant amount of data that could better uncover fraudulent activity that was simply not captured in 2008-2009. With the advent of “Big Data” and the hundreds or thousands of points of metadata that are routinely collected in 2023, I feel there may be more that could uncover fraudulent activity.

Another limitation of the research is how the target of “Potential Fraud” is based for the provider as a whole, and not at a transaction level. While this is rarely the case, where a provider commits fraud on every claim, the way this target variable is applied confounds both fraudulent and legitimate claims, making the modeling less accurate at both the transactional and provider levels. Because of this, provider information was later reattached to output results to give a ratio of total claims per provider, and total predicted fraud claims per provider. From there, a ratio of these two variables was calculated with the intent to group providers in risk categories where probabilities of fraud were occurring.

An additional limitation would be model tuning. For most models, there may be optimization that can be done to further improve the models with hyperparameter tuning.

# **Methods**

## *Data Collection and Processing*

The data used for this research was found at <https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis> and came in 8 files for various component of beneficiary data, inpatient and outpatient data, and whether there was probable fraud which was the target for modeling.

Once merged and ready for processing, the dataset contained 558,211 rows and 54 columns in the training data. For the test data, there were 135,392 rows and 53 columns since target of whether or not fraud may have occurred is excluded.

## *Data Dictionary*

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Description** | **Type of Data** |
| **Target** | | |
| PotentialFraud | Class label to indicate whether the Provider is a fraud or not | string |
| **Beneficiary Data** | | |
| DOB | Date of birth of the beneficiary | string |
| DOD | Date of death of the beneficiary in case the beneficiary is dead, otherwise, this contains the NaN value | string |
| Gender | Numeric code for gender | int64 |
| Race | Numeric code for race | int64 |
| RenalDiseaseIndicator | Flag to indicate whether the beneficiary has any problem related to kidney failure or not | string |
| State | Numeric code for state | int64 |
| County | Numeric code for county | int64 |
| NoOfMonths\_PartACov | Number of Months Patient has had Medicare Part A Coverage | int64 |
| NoOfMonths\_PartBCov | Number of Months Patient has had Medicare Part B Coverage | int64 |
| ChronicCond\_Alzheimer | Whether a beneficiary has a chronic disease | int64 |
| ChronicCond\_Heartfailure |
| ChronicCond\_KidneyDisease |
| ChronicCond\_Cancer |
| ChronicCond\_ObstrPulmonary |
| ChronicCond\_Depression |
| ChronicCond\_Diabetes |
| ChronicCond\_IschemicHeart |
| ChronicCond\_Osteoporasis |
| ChronicCond\_rheumatoidarthritis |
| ChronicCond\_stroke |
| IPAnnualReimbursementAmt | Annual amount reimbursed for the treatment of the beneficiary when admitted to the hospital | int64 |
| IPAnnualDeductibleAmt | Annual premium amount paid to the Insurance Agency towards the treatment of the beneficiary when admitted to the hospital | int64 |
| OPAnnualReimbursementAmt | Annual amount reimbursed for the treatment of the beneficiary when visited the hospital but not admitted | int64 |
| OPAnnualDeductibleAmt | Annual premium amount paid to the Insurance Agency towards the treatment of the beneficiary when he visited the hospital but was not admitted | int64 |
| **Inpatient and Outpatient Data** | | |
| BeneID | Unique identifier of the beneficiary who is registered to the health insurance provided by the Payer | string |
| ClaimID | Unique identifier of the health insurance claim | string |
| ClaimStartDt | Dates indicating when the insurance claim was submitted | string |
| ClaimEndDt | Dates indicating when the insurance claim was settled and closed | string |
| Provider | Unique identifier for a Provider (Hospital, Pharmacy, Laboratories, etc) | string |
| InscClaimAmtReimbursed | Amount reimbursed by the Payer (Insurance Agency) for the healthcare services provided to the beneficiary | int64 |
| AttendingPhysician | Columns showing the physicians who attended the beneficiary/patient, operated the patient, and any other physicians if any | string |
| OperatingPhysician | string |
| OtherPhysician | string |
| AdmissionDt | Columns showing the dates on which the beneficiary was admitted to the hospital and when he was discharged | string |
| DischargeDt | string |
| ClmAdmitDiagnosisCode | Diagnosis code indicating the beneficiary’s initial diagnosis at admission | string |
| DeductibleAmtPaid | Amount the beneficiary has to pay as part of the claim, and the rest of the amount is paid by the insurance company | float64 |
| DiagnosisGroupCode | Code to classify hospital cases, referred to as DRGs, which are expected to have similar hospital resource use (cost) | string |
| ClmDiagnosisCode\_1 | Diagnosis code identifying the beneficiary’s principal diagnosis | string |
| ClmDiagnosisCode\_2 | Diagnosis code in the nth position identifying the condition(s) for which the beneficiary is receiving care | string |
| ClmDiagnosisCode\_3 |
| ClmDiagnosisCode\_4 |
| ClmDiagnosisCode\_5 |
| ClmDiagnosisCode\_6 |
| ClmDiagnosisCode\_7 |
| ClmDiagnosisCode\_8 |
| ClmDiagnosisCode\_9 |
| ClmDiagnosisCode\_10 |
| ClmProcedureCode\_1 | Codes that indicate the principal or other procedures performed during the period covered by the institutional claim | float64 |
| ClmProcedureCode\_2 |
| ClmProcedureCode\_3 |
| ClmProcedureCode\_4 |
| ClmProcedureCode\_5 |
| ClmProcedureCode\_6 |

## *Univariate Statistics – Numeric Features*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Field Name** | **min** | **max** | **mean** | **std** | **NaN Count** |
| ClaimStartDt | NaN | NaN | NaN | NaN | 0 |
| ClaimEndDt | NaN | NaN | NaN | NaN | 0 |
| InscClaimAmtReimbursed | 0 | 125000 | 997.012133 | 3821.534891 | 0 |
| DeductibleAmtPaid | 0 | 1068 | 78.421085 | 274.016812 | 899 |
| AdmissionDt | NaN | NaN | NaN | NaN | 517737 |
| DischargeDt | NaN | NaN | NaN | NaN | 517737 |
| DOB | NaN | NaN | NaN | NaN | 0 |
| DOD | NaN | NaN | NaN | NaN | 554080 |
| IPAnnualReimbursementAmt | -8000 | 161470 | 5227.971466 | 11786.27473 | 0 |
| IPAnnualDeductibleAmt | 0 | 38272 | 568.756807 | 1179.172616 | 0 |
| OPAnnualReimbursementAmt | -70 | 102960 | 2278.225348 | 3881.846386 | 0 |
| OPAnnualDeductibleAmt | 0 | 13840 | 649.698745 | 1002.020811 | 0 |

## *Univariate Statistics – Categorical Features*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Field Name** | **min** | **max** | **mean** | **std** | **NaN Count** |
| BeneID | NaN | NaN | NaN | NaN | 0 |
| ClaimID | NaN | NaN | NaN | NaN | 0 |
| Provider | NaN | NaN | NaN | NaN | 0 |
| AttendingPhysician | NaN | NaN | NaN | NaN | 1508 |
| OperatingPhysician | NaN | NaN | NaN | NaN | 443764 |
| OtherPhysician | NaN | NaN | NaN | NaN | 358475 |
| ClmDiagnosisCode\_1 | NaN | NaN | NaN | NaN | 10453 |
| ClmDiagnosisCode\_2 | NaN | NaN | NaN | NaN | 195606 |
| ClmDiagnosisCode\_3 | NaN | NaN | NaN | NaN | 0 |
| ClmDiagnosisCode\_4 | NaN | NaN | NaN | NaN | 0 |
| ClmDiagnosisCode\_5 | NaN | NaN | NaN | NaN | 0 |
| ClmDiagnosisCode\_6 | NaN | NaN | NaN | NaN | 0 |
| ClmDiagnosisCode\_7 | NaN | NaN | NaN | NaN | 0 |
| ClmDiagnosisCode\_8 | NaN | NaN | NaN | NaN | 0 |
| ClmDiagnosisCode\_9 | NaN | NaN | NaN | NaN | 0 |
| ClmDiagnosisCode\_10 | NaN | NaN | NaN | NaN | 0 |
| ClmProcedureCode\_1 | 11 | 9999 | 5896.154612 | 3050.489933 | 0 |
| ClmProcedureCode\_2 | 42 | 9999 | 4106.358106 | 2031.640878 | 0 |
| ClmProcedureCode\_3 | 42 | 9999 | 4221.123839 | 2281.849885 | 0 |
| ClmProcedureCode\_4 | 42 | 9986 | 4070.262712 | 2037.62699 | 0 |
| ClmProcedureCode\_5 | 2724 | 9982 | 5269.444444 | 2780.071632 | 0 |
| ClmProcedureCode\_6 | NaN | NaN | NaN | NaN | 0 |
| ClmAdmitDiagnosisCode | NaN | NaN | NaN | NaN | 412312 |
| DiagnosisGroupCode | NaN | NaN | NaN | NaN | 517737 |
| Gender | 1 | 2 | 1.578838 | 0.493746 | 0 |
| Race | 1 | 5 | 1.255011 | 0.717437 | 0 |
| RenalDiseaseIndicator | NaN | NaN | NaN | NaN | 0 |
| State | 1 | 54 | 25.446969 | 15.192784 | 0 |
| County | 0 | 999 | 378.588195 | 265.215531 | 0 |
| NoOfMonths\_PartACov | 0 | 12 | 11.931472 | 0.889712 | 0 |
| NoOfMonths\_PartBCov | 0 | 12 | 11.93877 | 0.7859 | 0 |
| ChronicCond\_Alzheimer | 1 | 2 | 1.598132 | 0.490276 | 0 |
| ChronicCond\_Heartfailure | 1 | 2 | 1.409573 | 0.491755 | 0 |
| ChronicCond\_KidneyDisease | 1 | 2 | 1.587998 | 0.492196 | 0 |
| ChronicCond\_Cancer | 1 | 2 | 1.848615 | 0.358424 | 0 |
| ChronicCond\_ObstrPulmonary | 1 | 2 | 1.68707 | 0.463687 | 0 |
| ChronicCond\_Depression | 1 | 2 | 1.565193 | 0.495732 | 0 |
| ChronicCond\_Diabetes | 1 | 2 | 1.294605 | 0.455866 | 0 |
| ChronicCond\_IschemicHeart | 1 | 2 | 1.240735 | 0.42753 | 0 |
| ChronicCond\_Osteoporasis | 1 | 2 | 1.682353 | 0.465562 | 0 |
| ChronicCond\_rheumatoidarthritis | 1 | 2 | 1.688829 | 0.462973 | 0 |
| ChronicCond\_stroke | 1 | 2 | 1.89828 | 0.302279 | 0 |
| PotentialFraud | NaN | NaN | NaN | NaN | 0 |

## *Data Preparation*

Data was processed in several ways for the modeling that was later used. One hot encoding and min max scaling were two methods used. Scaling was handled differently for large value features versus other features in the dataset. Several features were removed from the dataset as they provided little data to move forward with where imputation would be difficult and likely misleading. There were several new features calculated that were later used including length of stay, length of claim processing, if a physician was listed in all three categories of attending, operating, and listed as other physician.

The nature of the data coming in multiple files required merging, which was not difficult to do. Each file was connected by Beneficiary ID and Claim ID, which were later dropped after merging the datasets. Imputation occurred for a numeric feature that I thought was important, which was Deductible amount paid. Here, I imputed using the mean of the feature to fill NaN values.

## *Feature Selection*

For feature selection, I found the following were strongest features using Top-K method:

|  |  |
| --- | --- |
| **Feature** | **Count** |
| Claim\_Period | 5 |
| InscClaimAmtReimbursed | 4 |
| OPAnnualReimbursementAmt | 4 |
| Patient\_Risk\_Score | 4 |
| Patient\_Age | 4 |
| Hospitalization\_Duration | 3 |
| IPAnnualReimbursementAmt | 3 |
| ChronicCond\_Alzheimer | 3 |
| ChronicCond\_Osteoporasis | 3 |
| ChronicCond\_rheumatoidarthritis | 3 |

## *Modeling*

For my modeling, I chose to do 5 different models; Logistic regression, Naïve bayes, Random Forest, Decision Tree and XGBoost. For XGBoost, I also tunes using grid search to find a more optimal set of parameters, but I found doing gave no increase in accuracy while drastically increasing the amount of computing power required to complete the model.

All models were output with model.predict() function, which predicts a 1 or 0 for any given row. While Logistic Regression output probabilistically using model.predict\_proba() method, the model.predict() method produced a 1 for positive prediction probability of 0.5 or greater, and produced a 0 for positive prediction probability under 0.5.

Both Decision Tree and Random Forest struck the best balance between Precision and Recall compared to others. These are also the two models that used several features to determine output, while the others keyed in heavily on hospitalization duration.

### Aggregate Modeling Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **1-F1 Bias** | **Variance** | **1-AUC-ROC** | **Variance** |
| Logistic Regression | 0.630 | 0.578 | 0.110 | 0.2692 | 0.0056 | 0.5388 | 0.0023 |
| Naïve Bayes | 0.619 | 0.502 | 0.149 | 0.3237 | 0.0027 | 0.5319 | 0.0026 |
| **Random Forest** | **0.655** | **0.561** | **0.438** | **0.7123** | **0.0019** | **0.7874** | **0.0013** |
| **Decision Tree** | **0.642** | **0.531** | **0.522** | **0.6551** | **0.0017** | **0.6605** | **0.0022** |
| XGBoost | 0.633 | 0.587 | 0.116 | 0.5676 | 0.003 | 0.7054 | 0.0023 |
| XGBoost Grid Search | 0.633 | 0.599 | 0.120 | 0.5797 | 0.0024 | 0.7237 | 0.0014 |

# **Citations**

Purnita, K. S., Mondal, M. R. H. (2021). Machine Learning For Dco-ofdm Based Lifi. PLoS ONE, 11(16), e0259955. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0259955>

Lin, S., Law, K., Yeh, Y., Wu, K., Lai, J., Lin, C., … & Kao, C. (2022). Applying Machine Learning To Carotid Sonographic Features For Recurrent Stroke In Patients With Acute Stroke. Front. Cardiovasc. Med., (9). <https://www.frontiersin.org/articles/10.3389/fcvm.2022.804410/full>

Johnson, J. M., Khoshgoftaar, T. M. (2019). Medicare Fraud Detection Using Neural Networks. J Big Data, 1(6). <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0225-0>

Hancock, J. T. I., Bauder, R. A., Wang, H., Khoshgoftaar, T. M. (2023). Explainable Machine Learning Models For Medicare Fraud Detection.. <https://www.researchsquare.com/article/rs-3076353/v1>

K, S. S., Ilango, V. (2019). A Hybrid Technique For Health Insurance Fraud Detection On Highly Imbalanced Dataset. IJITEE, 11(8), 3498-3501. <https://www.ijitee.org/wp-content/uploads/papers/v8i11/K24890981119.pdf>

Perols, J. (2011). Financial Statement Fraud Detection: An Analysis Of Statistical and Machine Learning Algorithms. AUDITING: A Journal of Practice &Amp; Theory, 2(30), 19-50. <https://publications.aaahq.org/ajpt/article-abstract/30/2/19/5673/Financial-Statement-Fraud-Detection-An-Analysis-of>

# **Appendix**

## *Data Pipeline and EDA*

<https://github.com/mulcahrj/6999_Capstone/blob/main/DATA6999_Final_Project_Data_Audit.ipynb>

## *Feature Selector*

<https://github.com/mulcahrj/6999_Capstone/blob/main/DATA6999_Feature_Selection.ipynb>

## *Links to Models*

### Naïve Bayes

<https://github.com/mulcahrj/6999_Capstone/blob/main/DATA6999_Naive_Bayes.ipynb>

### Logistic Regression

<https://github.com/mulcahrj/6999_Capstone/blob/main/DATA6999_Logistic_Regression.ipynb>

### Random Forest

<https://github.com/mulcahrj/6999_Capstone/blob/main/DATA6999_Random_Forest.ipynb>

### Decision Tree

<https://github.com/mulcahrj/6999_Capstone/blob/main/DATA6999_Decision_Tree.ipynb>

### XGBoost

<https://github.com/mulcahrj/6999_Capstone/blob/main/DATA6999_XGBoost.ipynb>

### XGBoost Grid Search

<https://github.com/mulcahrj/6999_Capstone/blob/main/DATA6999_XGBoost_Grid_Search.ipynb>