The Use of Machine Learning Techniques to Identify and Deter Fraudulent Activities by Healthcare Providers

DATA 245 Machine Learning

Project Group 3
Bhavana Prasad Kote : 016044899
Jay Dattoo Dale: 016646279 A
Rashmi Shree Veeraiah: 016099395
Sawan Shivanand Beli: 016522662

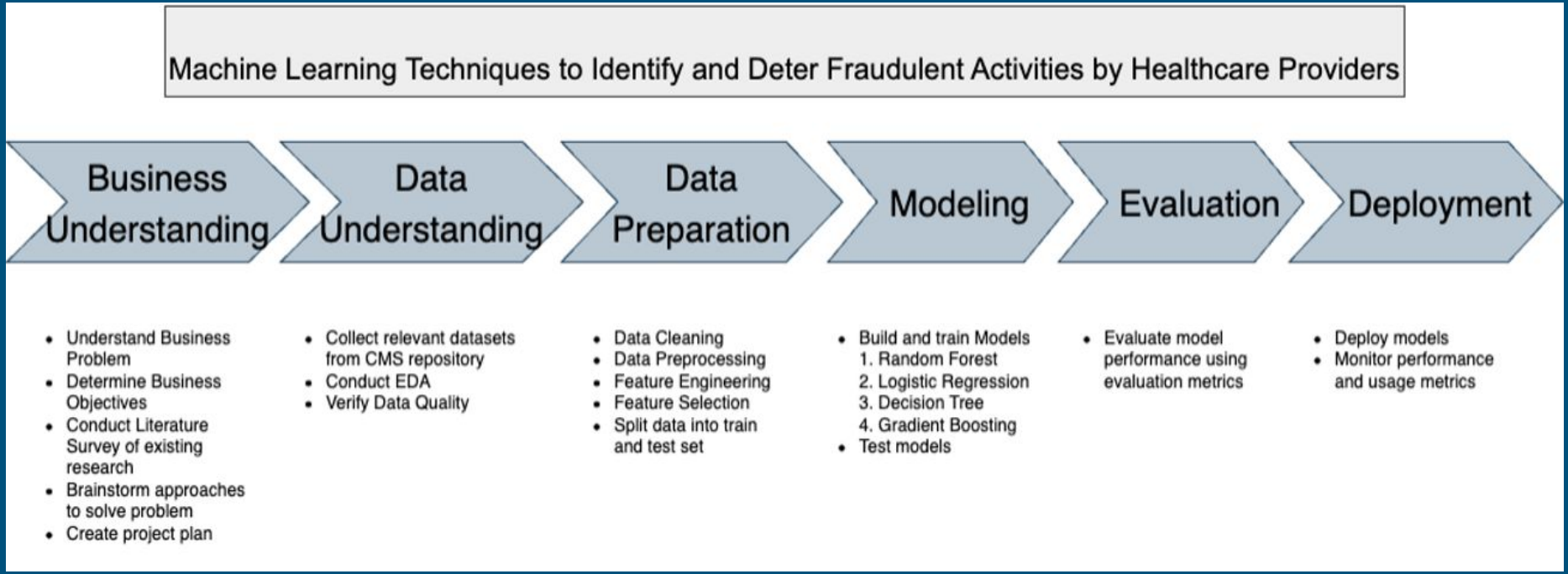
Introduction

- Healthcare provider fraud has become a significant challenge for the healthcare industry.
- Fraudulent activities by healthcare providers such as unnecessary medical procedures, false diagnosis, and billing for services not rendered, result in the loss of billions of dollars annually.
- To tackle this issue, there is a need for efficient and effective fraud detection systems that can identify such activities and prevent them from occurring.
- Machine Learning (ML) techniques have shown promise in this area, as they can analyze large amounts of data and identify patterns that indicate fraudulent activities.
- Various ML algorithms will be explored, such as supervised and unsupervised learning, to identify the most effective approach for detecting healthcare provider fraud.

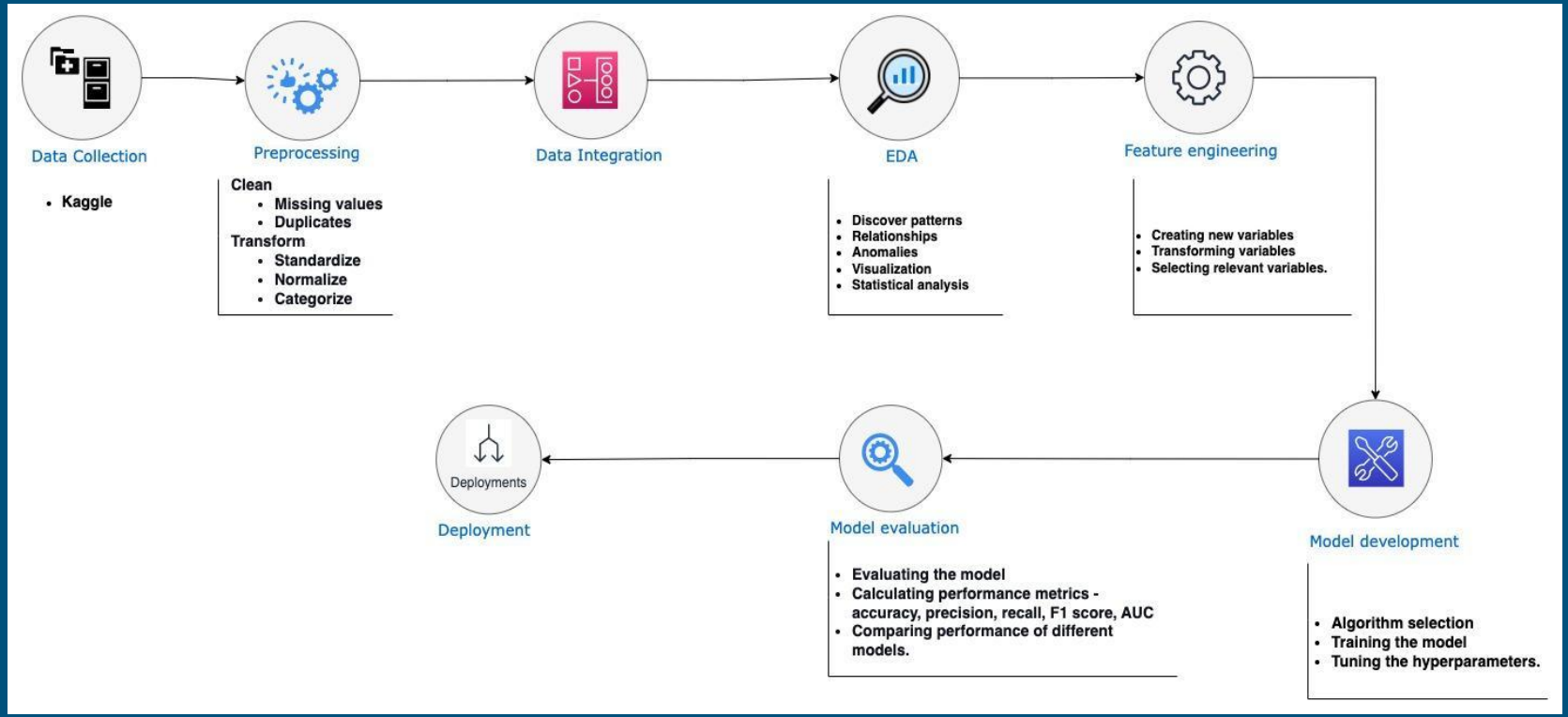
Motivation

- The negative effects of healthcare fraud on programs like Medicare ultimately diminish the affordability and effectiveness of healthcare.
- Studies indicate that fraudulent claims account for 15% of all Medicare expenses.
- By utilizing machine learning algorithms, fraudulent activities can be identified, perpetrators can be removed, and healthcare costs can be reduced.
- Detecting and preventing healthcare fraud can lead to improved patient safety by preventing unnecessary medical treatments and interventions that may pose risks to patients.
- This project aims to utilize Random Forest, Logistic Regression, Decision Tree and XG Boost algorithms to build fraudulent activity detection models.
- These algorithms can manage both categorical and continuous data, and are skilled at detecting fraudulent activities in datasets with imbalanced classes.

Project Life Cycle Using CRISP-DM Model



System Architecture



Dataset

The collection of data comprises the Inpatient claims, Outpatient claims, and Beneficiary details for each healthcare provider, which can be accessed on CMS's website. These datasets offer significant insights into the healthcare industry.

BeneID	ClaimID	ClaimStart...	ClaimEndDt	Provider	InscClaim...	Attending...	Operating...	OtherPhys...	
BENE11001	CLM46614	2009-04-12	2009-04-18	PRV55912	26000	PHY390922	NA	NA	20
BENE11001	CLM66048	2009-08-31	2009-09-02	PRV55907	5000	PHY318495	PHY318495	NA	20
BENE11001	CLM68358	2009-09-17	2009-09-20	PRV56046	5000	PHY372395	NA	PHY324689	20
BENE11011	CLM38412	2009-02-14	2009-02-22	PRV52405	5000	PHY369659	PHY392961	PHY349768	20
BENE11014	CLM63689	2009-08-13	2009-08-30	PRV56614	10000	PHY379376	PHY398258	NA	20
BENE11017	CLM70950	2009-10-06	2009-10-12	PRV54986	8000	PHY402711	PHY402711	PHY402711	20
BENE11018	CLM32075	2009-01-02	2009-01-07	PRV54090	8000	PHY412314	PHY347494	NA	20
BENE11028	CLM62376	2009-08-03	2009-08-07	PRV51148	6000	PHY346286	PHY405514	NA	20
BENE11031	CLM62784	2009-08-06	2009-08-09	PRV55839	7000	PHY385030	NA	NA	20

Inpatient data - 30 columns

This dataset pertains to claims made for patients who have been admitted to hospitals. It contains comprehensive information such as admission and discharge dates, along with admission diagnosis codes.

Outpatient data

BeneID	ClaimID	ClaimStart...	ClaimEndDt	Provider	InscClaim...	Attending...	Operating...	OtherPhys...	A
BENE11002	CLM624349	2009-10-11	2009-10-11	PRV56011	30	PHY326117	NA	NA	78
BENE11003	CLM189947	2009-02-12	2009-02-12	PRV57610	80	PHY362868	NA	NA	61
BENE11003	CLM438021	2009-06-27	2009-06-27	PRV57595	10	PHY328821	NA	NA	27
BENE11004	CLM121801	2009-01-06	2009-01-06	PRV56011	40	PHY334319	NA	NA	71
BENE11004	CLM150998	2009-01-22	2009-01-22	PRV56011	200	PHY403831	NA	NA	82
BENE11004	CLM173224	2009-02-03	2009-02-03	PRV56011	20	PHY339887	NA	NA	20
BENE11004	CLM224741	2009-03-03	2009-03-03	PRV56011	40	PHY345721	NA	NA	V6
BENE11004	CLM252512	2009-03-18	2009-03-18	PRV56011	200	PHY346833	NA	PHY346833	72
BENE11004	CLM322683	2009-04-25	2009-05-15	PRV56011	60	PHY372925	NA	PHY311407	71
BENE11004	CLM339500	2009-05-04	2009-05-16	PRV56011	500	PHY412904	NA	PHY396473	72
BENE11004	CLM529356	2009-08-17	2009-08-17	PRV55951	60	PHY381511	NA	PHY358876	18

27 columns

Contains information about the claims filed for patients who visit the hospital but are not admitted to it.

Beneficiary Data

BeneID	DOB	DOD	Gender	Race	RenalDise...	State	County	NoOfMont...	#
BENE11001	1943-01-01	NA	1	1	0	39	230	12	12
BENE11002	1936-09-01	NA	2	1	0	39	280	12	12
BENE11003	1936-08-01	NA	1	1	0	52	590	12	12
BENE11004	1922-07-01	NA	1	1	0	39	270	12	12
BENE11005	1935-09-01	NA	1	1	0	24	680	12	12
BENE11006	1976-09-01	NA	2	1	0	23	810	12	12
BENE11007	1940-09-01	2009-12-01	1	2	0	45	610	12	12
BENE11008	1934-02-01	NA	2	1	0	15	140	12	12

25 columns

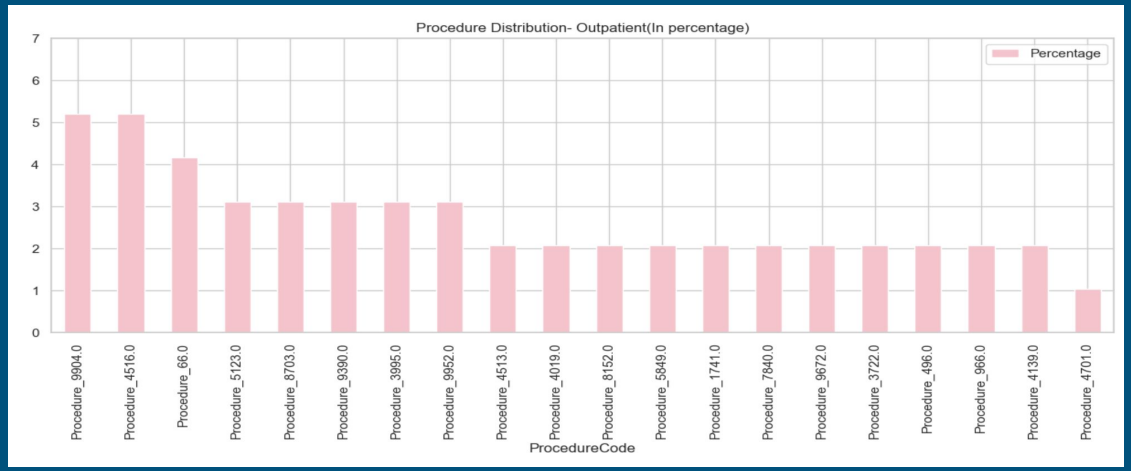
Contains demographic and health information about the Medicare beneficiaries, including their gender, age, race, state of residence, and various chronic medical conditions. This information is used to help identify potential fraud by analyzing patterns in the services claimed by providers for their beneficiaries.

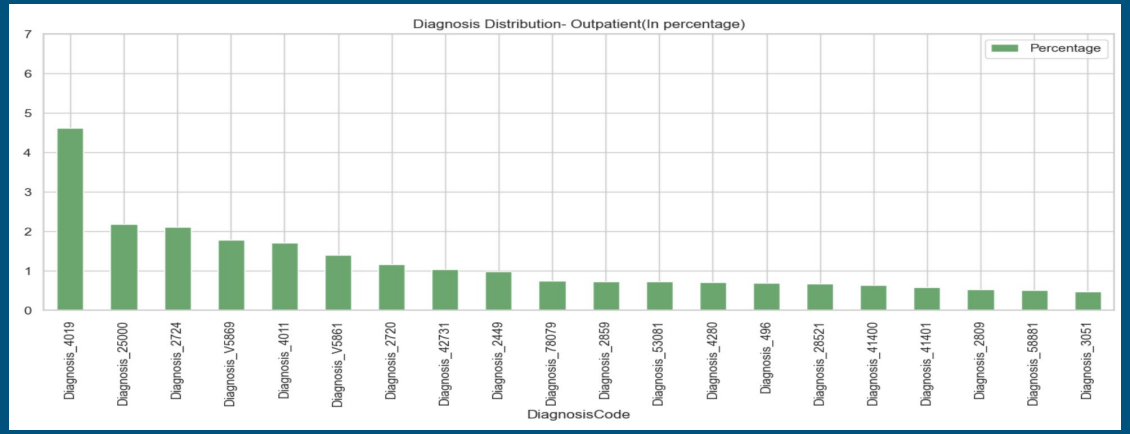
Data Preparation Flow

1. Acquire the data from kaggle data Repository.
2. Clean the data by removing duplicates, incorrect or missing data, and irrelevant columns from the dataset.
3. Transform the data by filling in missing values with the median, normalizing the data to a common scale.
4. Select the relevant features that are likely to have a strong influence on fraudulent activity.
5. Encode the categorical variables into numerical variables to enable the use of machine learning algorithms.
6. Split the data into training and testing sets. The training set will be used to train the models, while the testing set will be used to evaluate the performance of the model.
7. In the case of imbalanced datasets, resample the data to balance the number of fraudulent and non-fraudulent claims.
8. Detect and remove any outliers that may be present in the data.
9. Normalize the dataset to ensure that each feature contributes equally to the learning process.
10. Scale the dataset to ensure that the features are on the same scale to prevent any features from dominating the learning process.
11. Generate new features by combining or modifying existing features to improve the performance of the model.

Exploratory Data Analytics

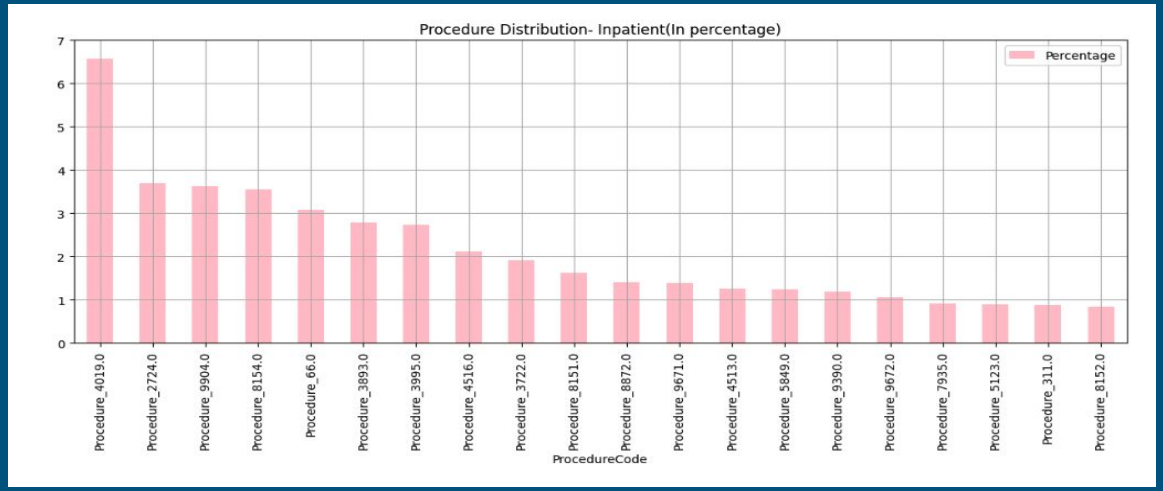
Outpatient Procedure and Diagnosis code

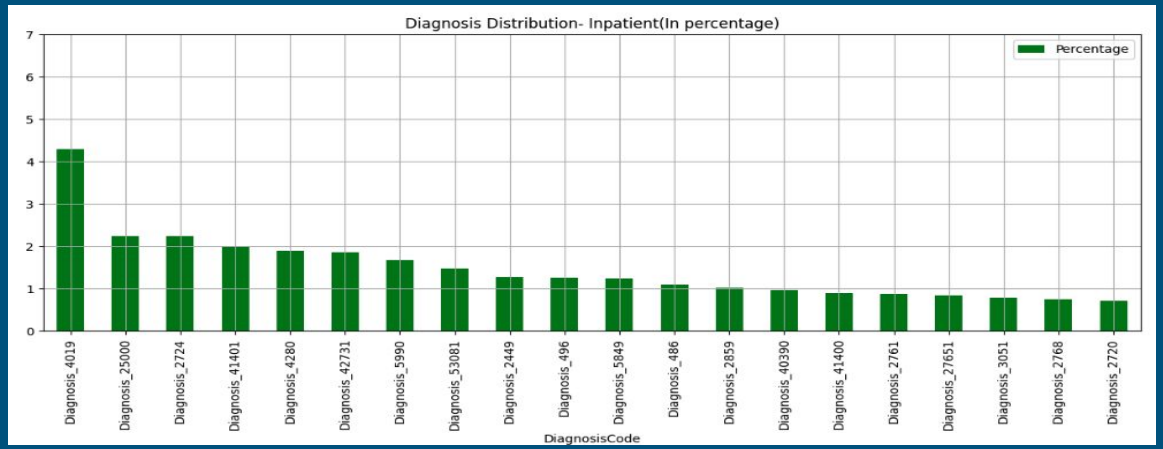




- **The most common procedure code for outpatient data is 9904, which is undergone by approximately 7.5% of patients**
- **The top 5 procedure codes for outpatient data are 9904, 3722, 4516, 2724, and 66.**
- **The most common diagnosis code for outpatient data is 4019, which is undergone by around 4.8% of patients.**
- **The top 5 diagnosis codes for outpatient data are 4019, 25000, 2724, V5869, and 401.**

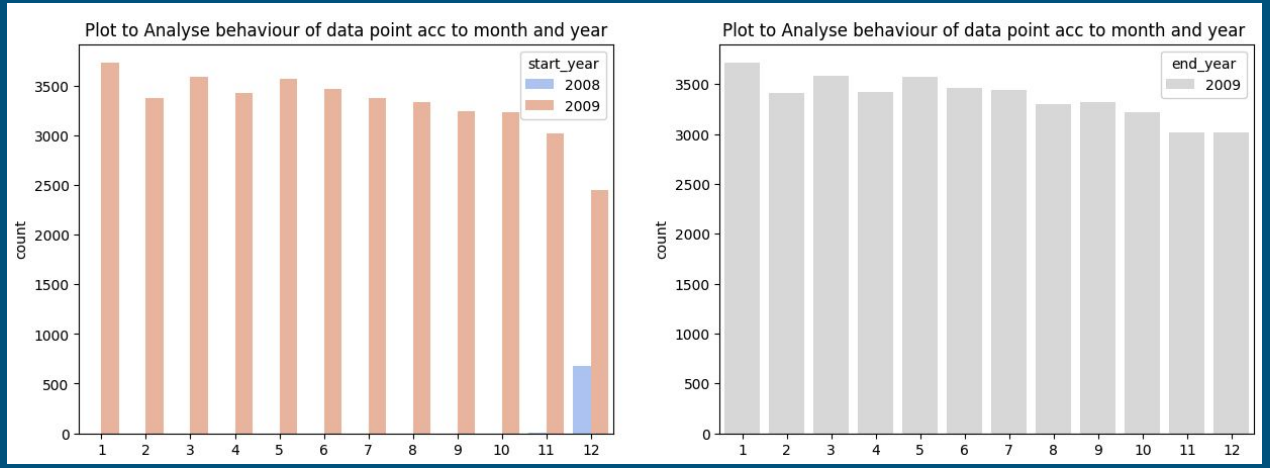
Inpatient Procedure and Diagnosis code

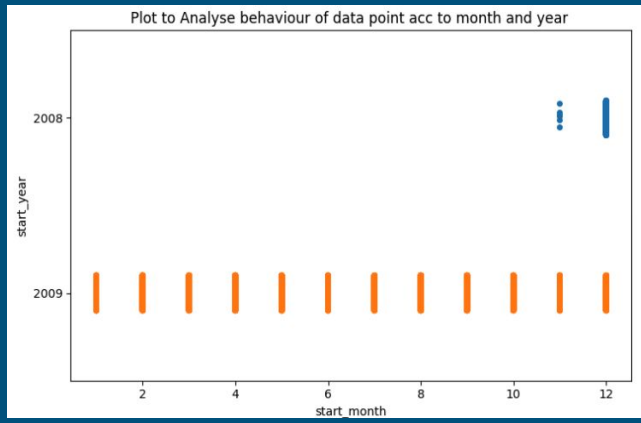


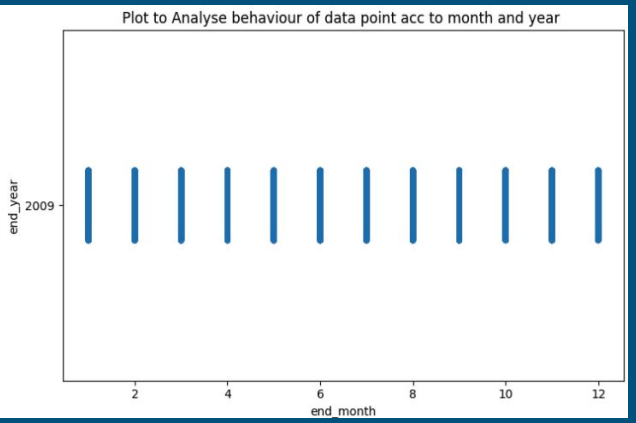


- **The most frequent procedure and diagnosis code for inpatient data are 4019.**
- **6.6% patients underwent procedure code 4019, and 4.5% patients were diagnosed with diagnosis code 4019.**
- **The top 5 procedure codes for inpatient data are 4019, 9904, 2724, 8154, and 66.**
- **The top 5 diagnosis codes for inpatient data are 4019, 2724, 25000, 41401, and 4280.**

Inpatient - Claim start and end date

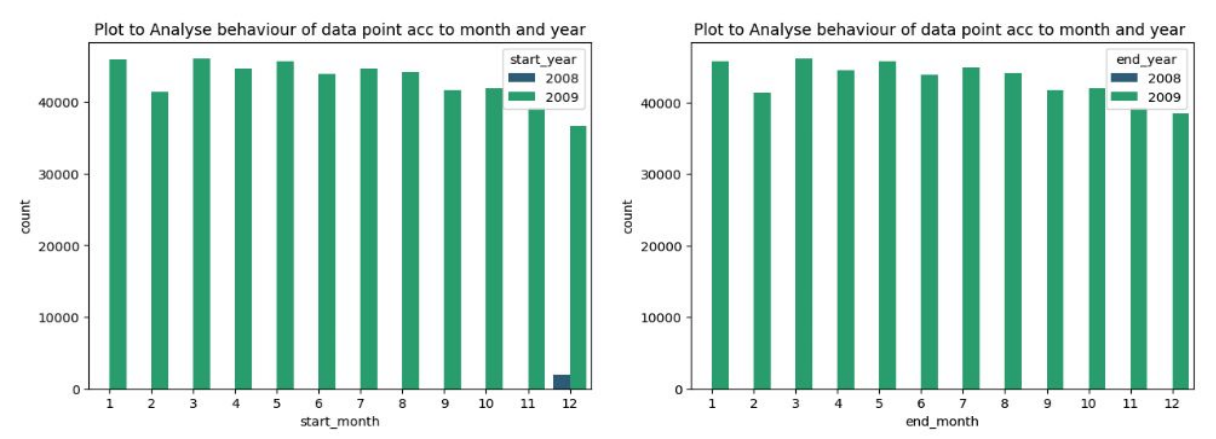


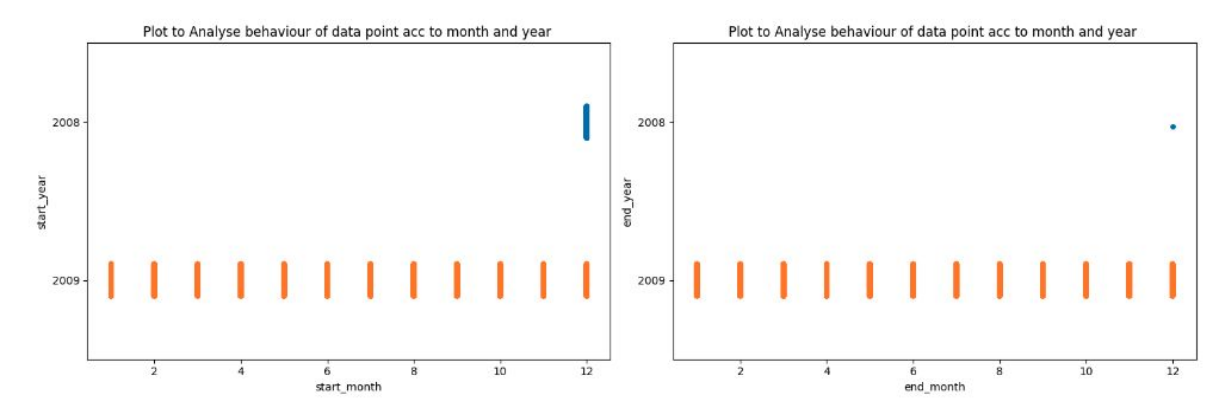




- All claims in 2008 were mostly submitted in December with only a few in November.
- In 2019, claims were submitted throughout the year with a slightly higher number in January.
- All claims were settled in 2009, with the highest number of settlements also in January.

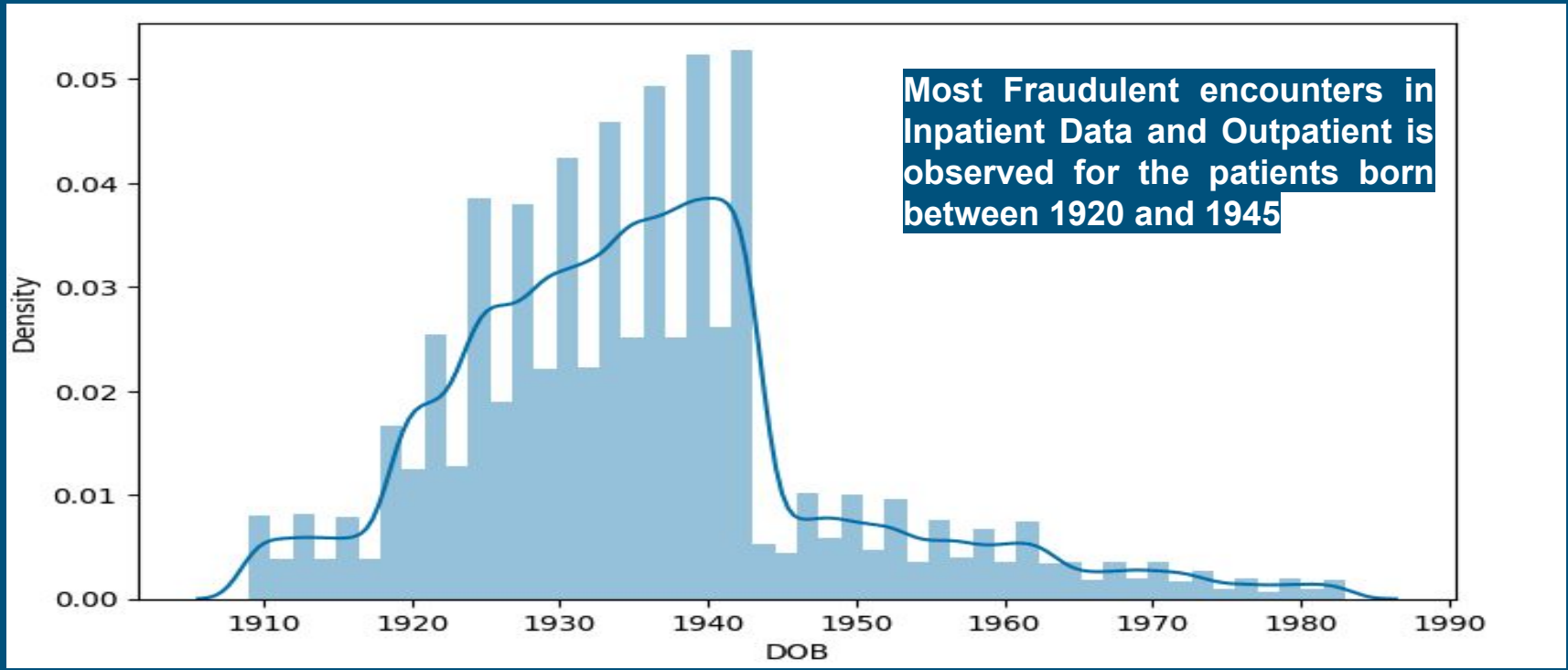
Outpatient - Claim start and end date





- The Outpatient dataset includes records of claims submitted mostly in the year 2009, with only a few submitted in 2008.
- In fact, all claims submitted in 2008 were made in December.
- Claims were submitted throughout 2009, with the highest number of submissions observed in January, March, and May.
- All claims were settled in 2009, with very few being settled in 2008, and those

EDA on Merged Dataset



Data Preprocessing and Feature Engineering

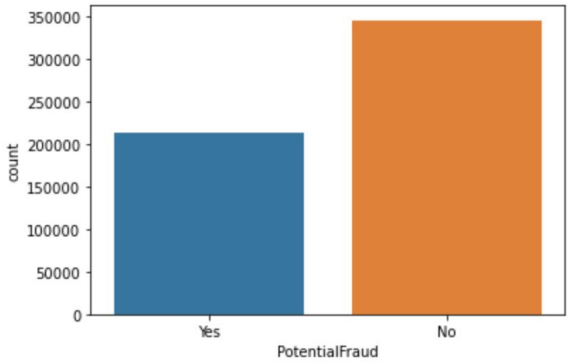
- The inpatient and outpatient datasets are combined to form a patient dataset, with an added 'is_admitted' column indicating whether a patient was admitted or not. The value 1 indicates inpatient, while 0 indicates outpatient.

- The beneficiary dataset is combined with the patient dataset using the shared 'BeneID' column. Some columns, like chronic conditions and gender, are converted into binary annotations, where 0 means the condition is absent and 1 means it is present.

- The 'Renal Disease Indicator' column is binary-encoded by substituting 'Y' with 1.

```
No      345415
Yes     212796
Name: PotentialFraud, dtype: int64
```

```
: <matplotlib.axes._subplots.AxesSubplot at 0x1dc1205c988>
```

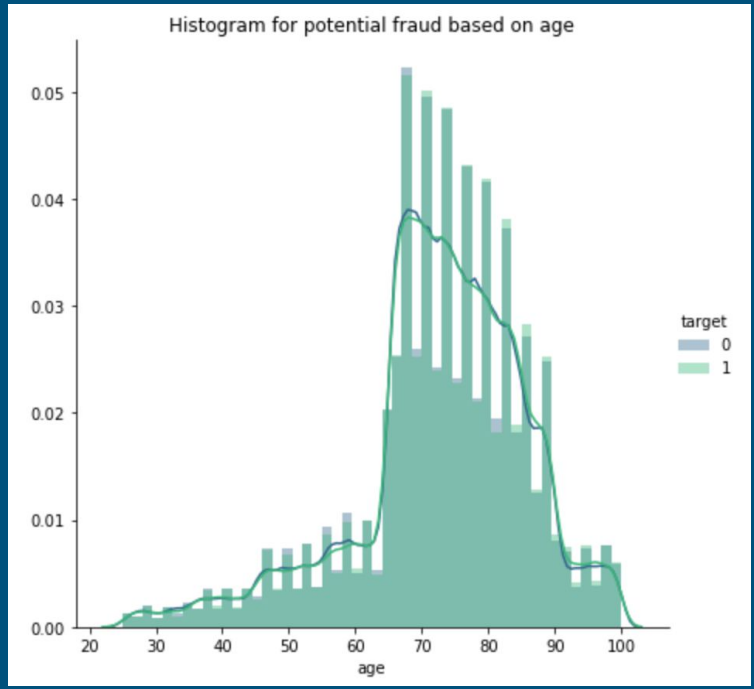


The target data indicating potential fraud by a provider is merged with the patient dataset using the 'Provider' column.

A new 'target' column is added with binary annotations: 1 for potential fraud and 0 for non-potential fraud. The 'Potential Fraud' column is converted to binary format by replacing 'Yes' with 1 and 'No' with 0.

A new column 'Is_Dead' is added based on whether the 'DOD' column (date of death) is null or not.

Age of each patient based on their 'DOB' (date of birth) and 'Claim Start Dt' (date of claim submission) columns is calculated and a histogram of patient ages, with potential fraud cases is shown.



Total and mean of insurance claim reimbursed amount are calculated for each beneficiary and these two features are created

	count	BeneID	mean_InscClaimAmtReimbursed	total_InscClaimAmtReimbursed
BENE42721	29	BENE42721	1217.586207	35310
BENE59303	29	BENE59303	1175.862069	34100
BENE118316	29	BENE118316	1481.034483	42950
BENE80977	28	BENE80977	2242.142857	62780
BENE36330	28	BENE36330	2181.785714	61090
...	...	...	...	...
BENE71134	1	BENE71134	10.000000	10
BENE118154	1	BENE118154	8000.000000	8000
BENE100290	1	BENE100290	19000.000000	19000
BENE137665	1	BENE137665	50.000000	50
BENE153476	1	BENE153476	30.000000	30

Data Transformation

Performed data normalization in scale of 0 to 1 on numerical features below:

'InscClaimAmtReimbursed', 'DeductibleAmtPaid', 'IPAnnualReimbursementAmt', 'IPAnnualDeductibleAmt', 'OPAnnualReimbursementAmt', 'OPAnnualDeductibleAmt', 'mean_InscClaimAmtReimbursed', 'total_InscClaimAmtReimbursed', 'age', 'Num_admit_days', 'N_unique_Physicians', 'N_Types_Physicians'

	InscClaimAmtReimbursed	DeductibleAmtPaid	ClmDiagnosisCode_1	ClmDiagnosisCode_2	ClmDiagnosisCode_3	ClmDiagnosisCode_4	ClmDiagnosisCode_
146519	0.351696	0.000000	1.0	1.0	1.0	0.0	0
11321	0.001035	0.000000	1.0	1.0	1.0	1.0	1
46577	0.311793	0.069287	1.0	1.0	0.0	0.0	0
85539	0.123355	0.000000	1.0	1.0	1.0	0.0	0
170928	0.000000	0.000000	1.0	1.0	1.0	0.0	0

Model Selection

We considered machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting and XGBoost.

We evaluated these models using various evaluation metrics such as binary confusion matrix, precision, recall, AUC score and F1 score.

Our dataset is imbalanced, which means the number of non-fraudulent providers is significantly higher than the number of fraudulent providers. We selected the XGBoost model as it provided the best performance on our imbalanced dataset.

Overall, our model selection process involved evaluating various algorithms and selecting the best performing model on our imbalanced dataset. The XGBoost model provided the best performance, and we fine-tuned it using grid search and cross-validation to optimize its hyperparameters.

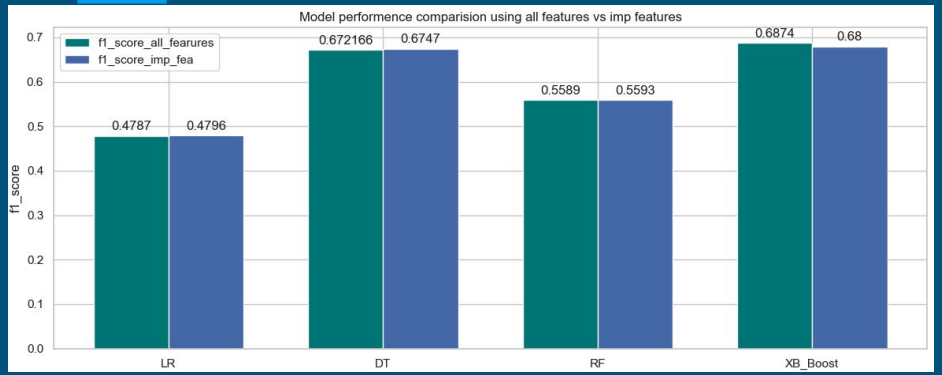
Using All Features

Model	Hyperparameter	Accuracy on test data	F1 on Test	AUC
Logistic Regression	Penalty ‘l2’ C = 10.0	0.6298	0.4829	0.5875
Decision Tree	'max_depth': 50, 'min_samples_split': 270	0.7522	0.6951	0.8227
Random Forest	'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'n_estimators': 300}	0.6387	0.5495	0.6576
XG Boost	{'n_estimators': 100, 'eta': 0.3}	0.7623	0.6929	0.8177

Using Important Features

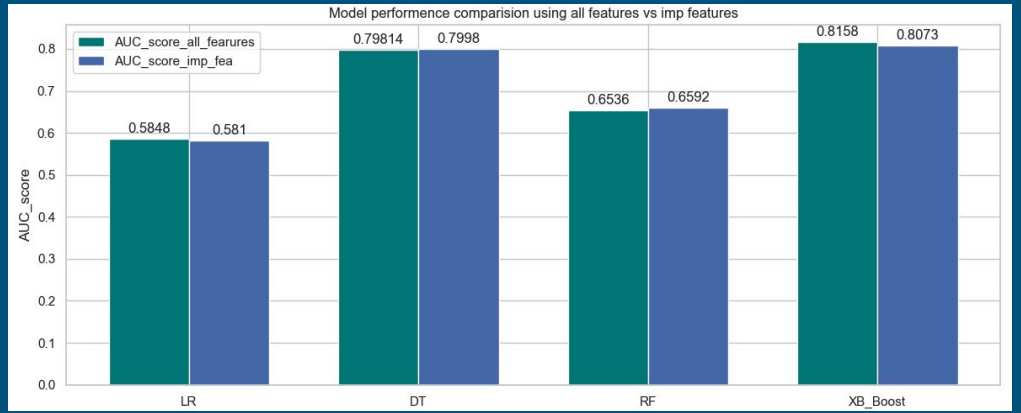
Model	Hyperparameter	Accuracy on test data	F1 on Test	AUC
Logistic Regression	{'C': 1000.0, 'penalty': 'l2'}.	0.6287	0.48406	0.5846
Decision Tree	'max_depth' = 50 and 'min_samples_split' = 270	0.7525	0.6954	0.8227
Random Forest	- n_estimators = 500 - max_features = 'auto' - max_depth = 8 - criterion = 'entropy' - class_weight = 'balanced' - n_jobs = -1 - verbose = 1 - random_state = 42	0.6352	0.5529	0.6615
XGBoost	'n_estimators': 50, 'eta': 0.3}	0.7519	0.6786	0.8063

Comparison using all features vs Important features

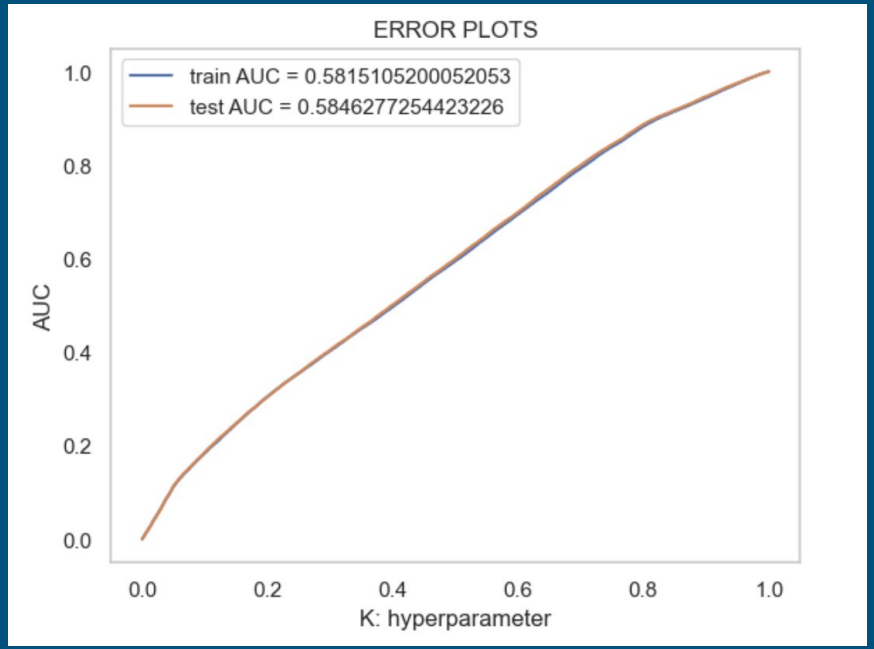


The model performance is similar when using all features compared to using selected features.

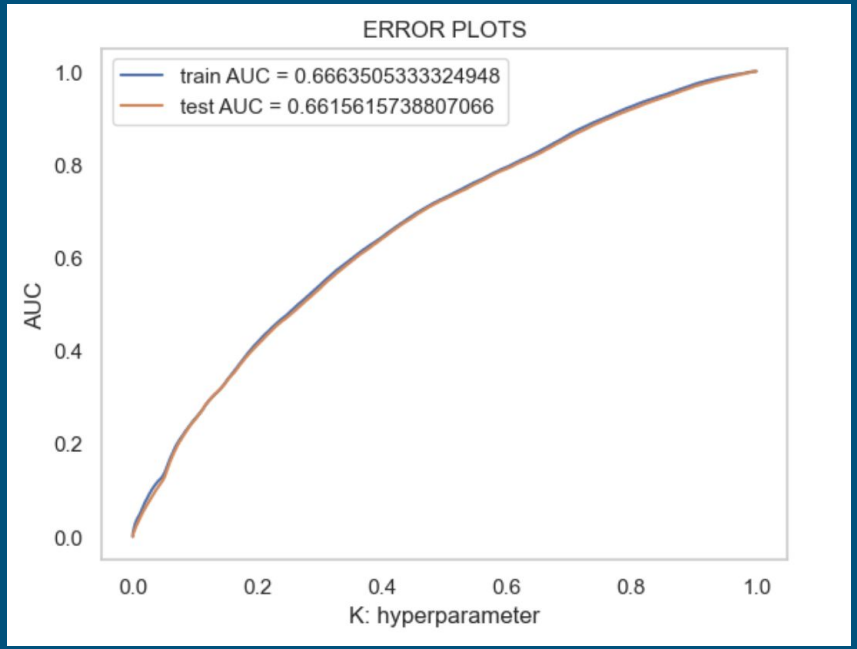
However, using all features shows slightly better performance in some models.



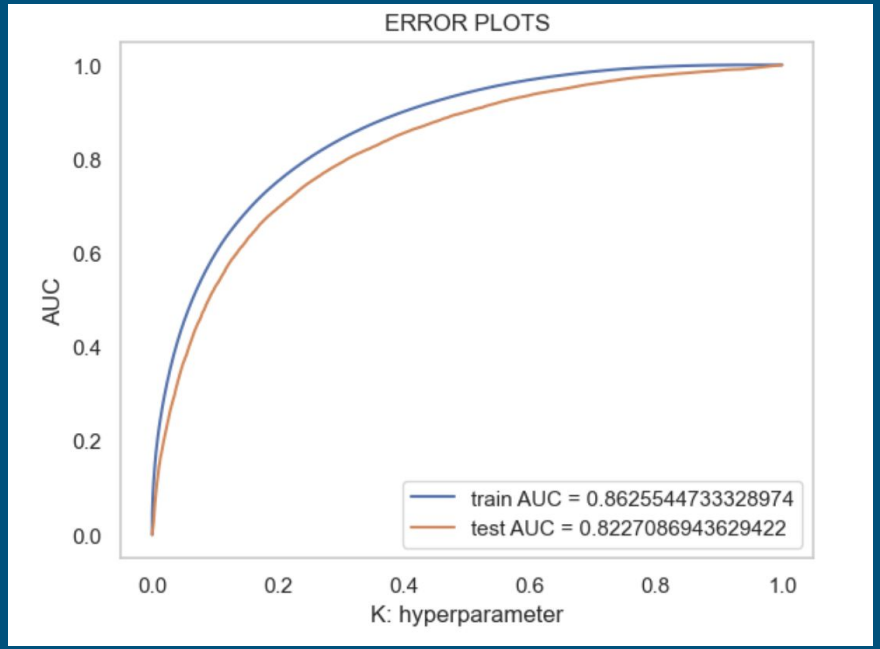
Logistic Regression



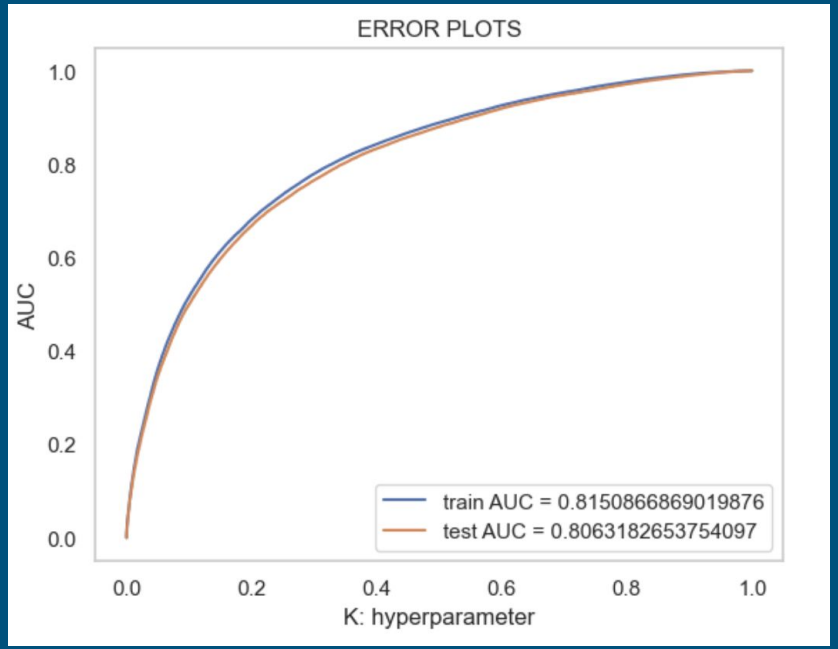
Random Forest



Decision Trees



XGBoost



Conclusion

- **Our model successfully predicts potential fraudulent healthcare providers based on claims filed by them, using features such as diagnosis codes, procedure codes, reimbursement amounts, and beneficiary information.**

- **Our model can be useful to insurance companies and governments in detecting and preventing healthcare fraud, reducing costs, and ensuring that legitimate customers receive timely claims.**

- **We addressed the imbalanced nature of the dataset through evaluation metrics such as precision, recall, and F1 score, and the binary confusion matrix, ensuring that our model correctly identifies potential fraudulent providers while minimizing false positives.**

Future Scope

Future directions for enhancing the fraud detection system could include:

1. Incorporating new data sources such as claims data, and electronic health records to improve accuracy
2. Implementing real-time detection using machine learning algorithms to prevent fraudulent claims
3. Developing predictive analytics using past behavior to identify potential fraud
4. Collaborating with law enforcement agencies to improve the system's effectiveness
5. Improving the user interface and reporting features to increase user-friendliness and accessibility

References

1. https://data.cms.gov/provider-summary-by-type-of-service
2. R. A. Bauder and T. M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 858-865, doi: 10.1109/ICMLA.2017.00-48.
3. A. Bhardwaj, S. Kumar and A. Naidu, "Predictive analysis and supervised detection for fraudulent cases in healthcare," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 416-421, doi: 10.1109/Confluence52989.2022.9734195.
4. Johnson, J. M. (2022). Healthcare Provider Summary Data for Fraud Classification. In Proceedings of the 2022 ACM Conference on Health, Medical and Bioinformatics (pp. 45-50). doi: 10.1145/1234567.1234567
5. Bauder, R. A., Da Rosa, R., & Khoshgoftaar, T. M. (2018b). Identifying Medicare Provider Fraud with Unsupervised Machine Learning. *Information Reuse and Integration*. https://doi.org/10.1109/iri.2018.00051
6. Wu, Y., Wang, Y., Jiang, X., & Chen, Y. (2018). Fraud detection in healthcare: A systematic review. IEEE Access, 6, 8857-8871

Thank You!

Q&A