Medicare Fraud Detection

Rishabh Sareen

Department of Mathematics

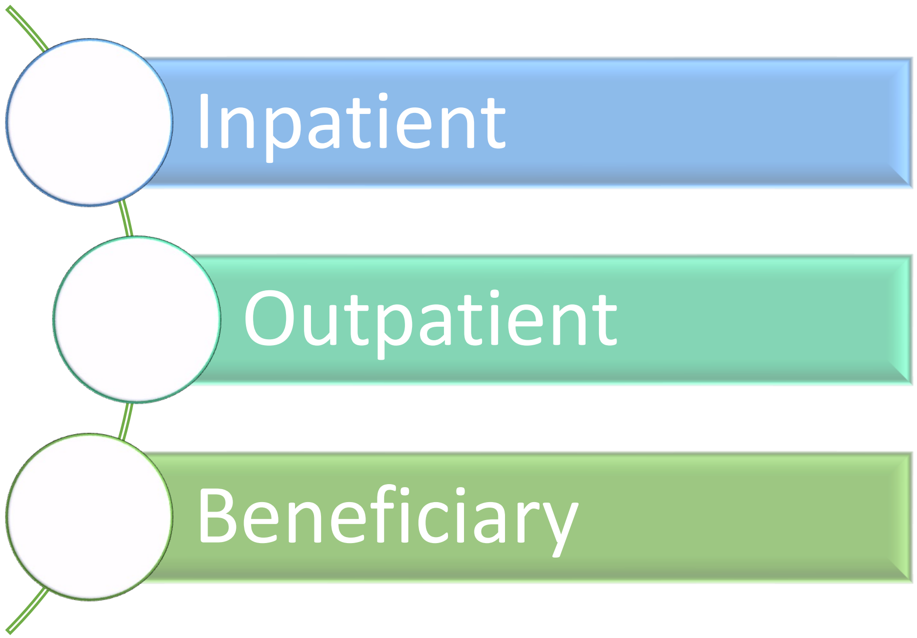
Michigan State University

What is Medicare fraud?

- Medicare Frauds are committed by Doctors/Hospitals aided by medical malpractices.
- Medical Providers try to maximize reimbursement received from Insurance companies via illegitimate activities such as submitting false claims:
- How do they commit fraud?
  - Billing for care that they never rendered.
  - Submitting duplicate claims.
  - Falsifying claim/patient info.
  - Disguising non-covered services as covered services.
- Objective: Build an innovative machine learning model that predicts fraud in the Medicare industry using anomaly analysis and geo-demographic metrics.

Medicare claims dataset

- ❑ Beneficiary:
  - Info about patients for whom claims have been submitted
- ❑ Inpatient:
  - Claim level data (with provider/doctor info) for the patients that have stayed at the hospital for the medical service.
- ❑ Outpatient:
  - Claim level data (with provider/doctor info) for the patients that have stayed at the hospital for the medical service.



Data Preprocessing

- ❑ Missing Data Imputed:
  - Beneficiary:
  - Inpatient:
  - Outpatient:
- ❑ Label Encoding: All categorical features
- ❑ New Feature Creation:
  - Deceased, Tot_Reimbursed_Amt, Hospital_Stay, Claim_Duration, Physician_Count, Claim_Count, Chr_Cond_Count, etc.
- ❑ Dropped features:
  - With high null values, ones from which other features were created, etc.
- ❑ Combined all datasets with fraud labels.

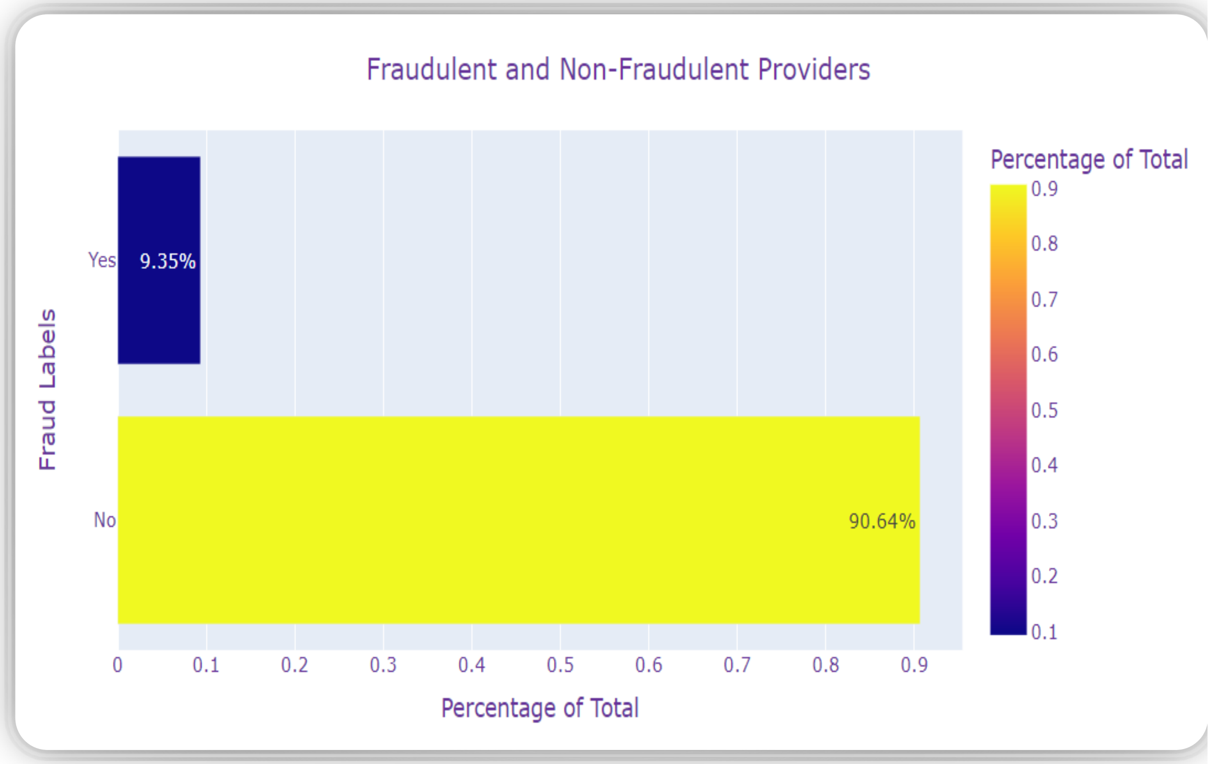
- ❑ Different model types attempted:
  - Logistic Regression
  - Random Forest
  - Linear SVC

  Hyperparameters were tuned based on the F1 metric. Final parameters are chosen by Recursive Feature Selection

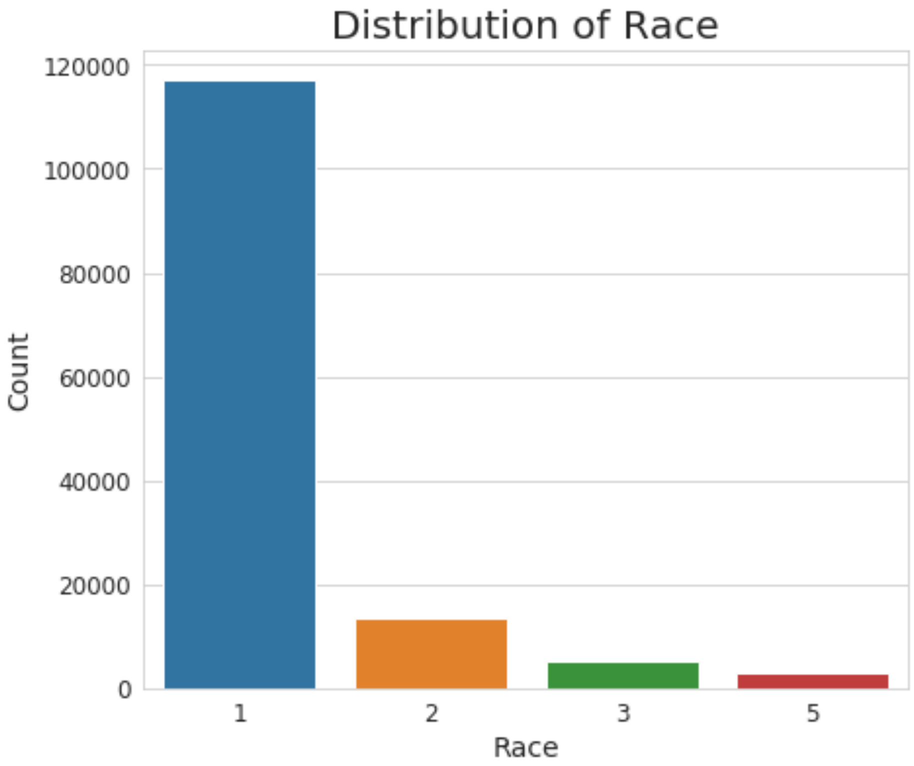
Missing Values	• Beneficiary: Date of Death • Inpatient: Diagnosis/Procedure codes • Outpatient: Operating physicians
Label Encoding	• All categorical features
Data Scaling	• Robust-Scaled due to presence of outliers
Upsampling	• Smote • BorderlineSmote

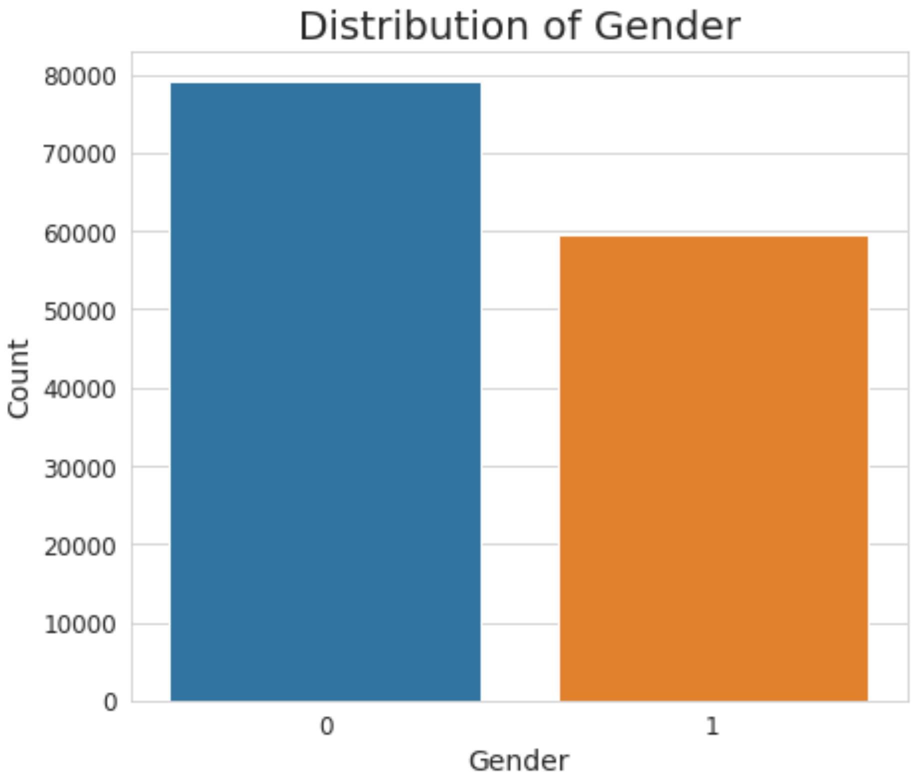
Fraud labels

- Fraud labels provided for Hospitals.
- Fraud/Non-Fraud providers:



Basic Patient Information:

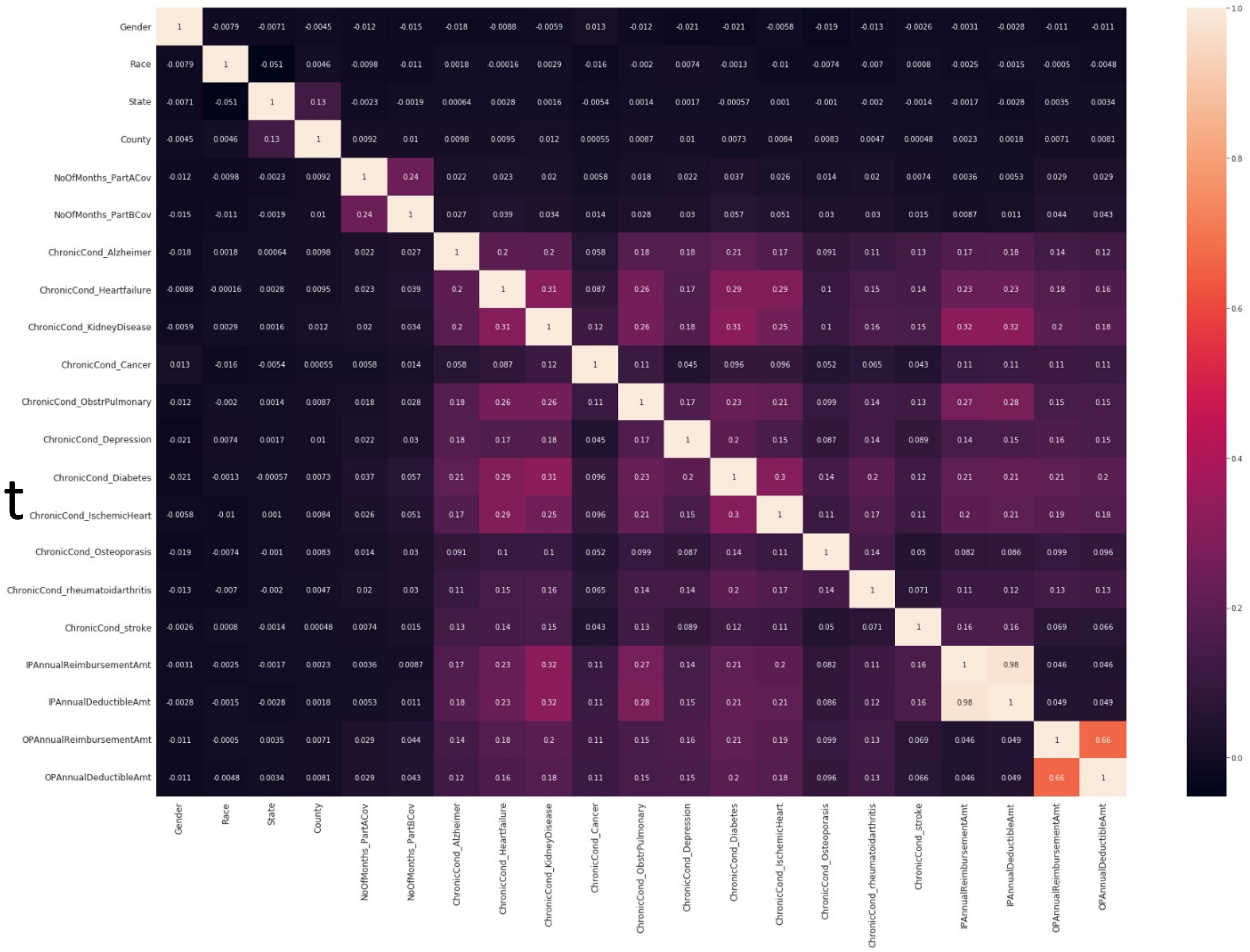




Correlation

- IPAnnualDeductibleAmt is collinear with IPAnnualReimbursementAmt 0.97.

- OPAnnualReimbursementAmt is collinear with OPAnnualDeductibleAmt 0.66.



Feature Selection

Feature Engineering

Feature Creation

- Deceased
- Hospital Stay
- Claim Duration
- Physician Count
- Claim Count
- Chronic Condition Counts

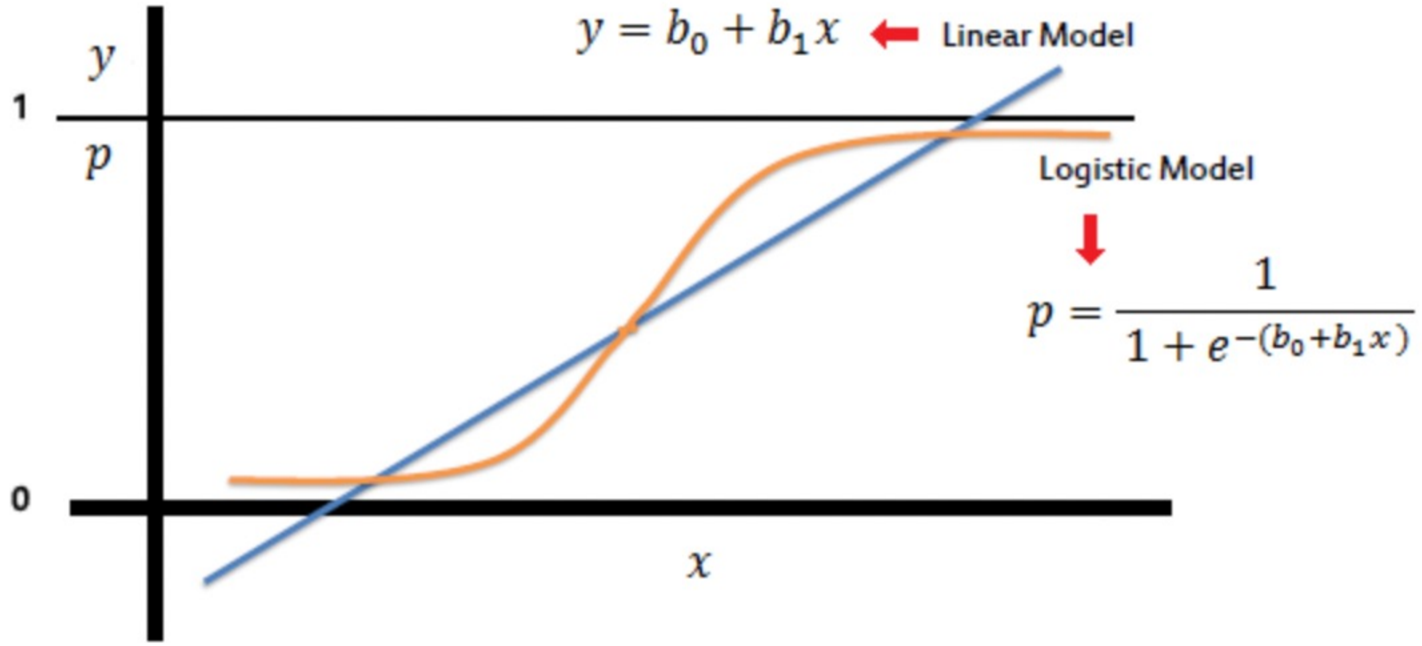
Dropped Features

- Features with high null values.
- Features from which other features were created.

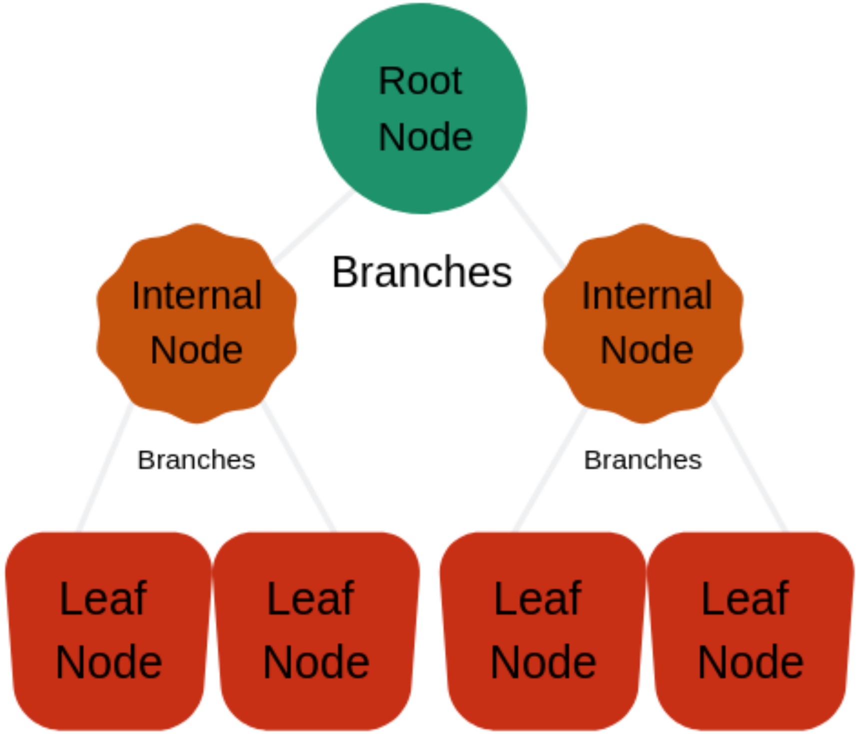
```
Features_pd1.count()

npi                        918012
total_drug_cost_sum        918012
total_drug_cost_mean       918012
total_drug_cost_max        918012
total_claim_count_sum      918012
total_claim_count_mean     918012
total_claim_count_max      918012
total_day_supply_sum       918012
total_day_supply_mean      918012
total_day_supply_max       918012
city                       918012
state                      918012
last_name                  917986
first_name                 918000
Speciality                 918012
Total_Payment_Sum          382724
is_fraud                      368
dtype: int64
```

Logistic Regression



Random Forest

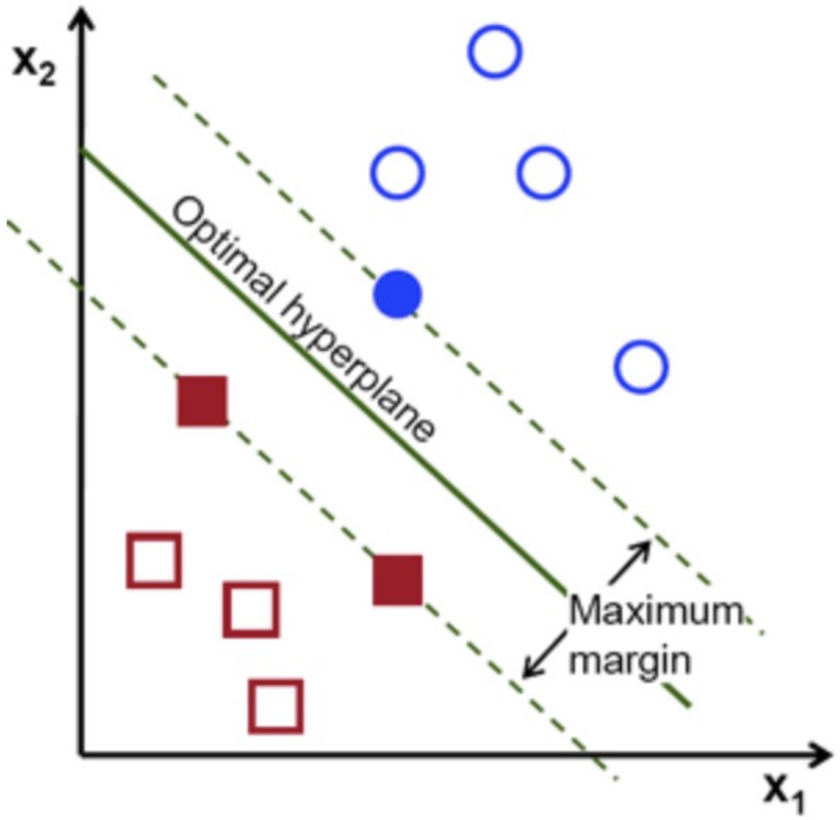


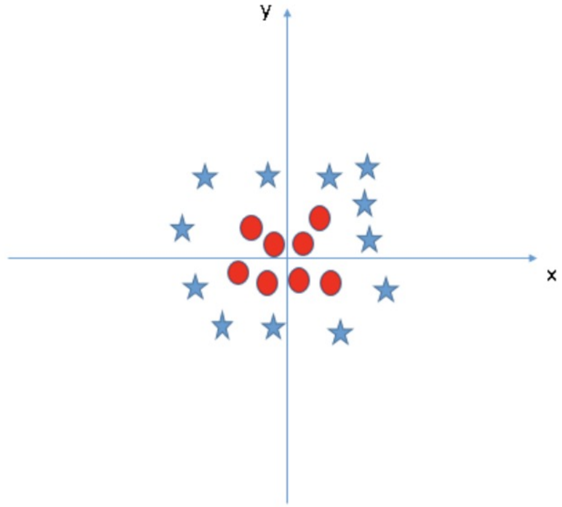
$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

Where *N* is the number of data points, *fi* is the value returned by the model and *yi* is the actual value for data point *i*.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Support Vector Algorithm





Gaussian Kernel

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

Results and Discussions

Sampling Ratio	Model	Features	AUC	F1 Score
80:20	Logistic Reg	All Features	0.951	0.551
80:20	Logistic Reg	Important Features	0.942	0.56
80:20	Random Forest	All Features	0.943	0.62
80:20	Random Forest	Important Features	0.943	0.634

Though F1 score is higher for Random forest indicating a higher accuracy of around 0.63 but the confusion matrix and AUC curve better results for Logistic Regression. Depending on the models employed for feature Selection, one can make a choice between all three models.

Conclusions

Possibly Fraud Providers	**Non-Fraud Providers**
High average claim reimbursement amounts. Some of these providers have the highest reimbursement amounts in the dataset.	Low average claim reimbursement amounts
High average number of patient insurance claims.	Low average number of patient insurance claims.
A narrow range of patient age.	A wider range of patient age.
A narrow range of total patient chronic condition counts.	A wider range of total patient chronic condition counts.
Outpatient – high number of diagnosis codes listed on claims	Outpatient - low number of diagnosis codes listed on claims

Conclusions

- Beneficiaries having high reimbursement amounts or paying high deductibles also have more chronic conditions and could be more susceptible to fraud.

- A patient’s age being in a certain range, and who their primary doctor is could in certain cases make them more vulnerable to fraud.