

Solving Class Imbalance Problem Using Inverse Random Under-Sampling (IRUS)

Group Members:

Minhaj Uddin Meraj 18K-0177

Zain Shaikh 18K-0331

Hamza Edhi 18K-0335



Identifying the Research Goal

- Our aim is to solve Class Imbalance Problem
- Our data set contains data very high number of samples in one class and very few samples in other class (class imbalance)
- To tackle this problem we carried out the research and thus used a novel approach known as **Inverse Random Under Sampling (IRUS)**
- This method has been used previously to solve class imbalance problem for binary classes.

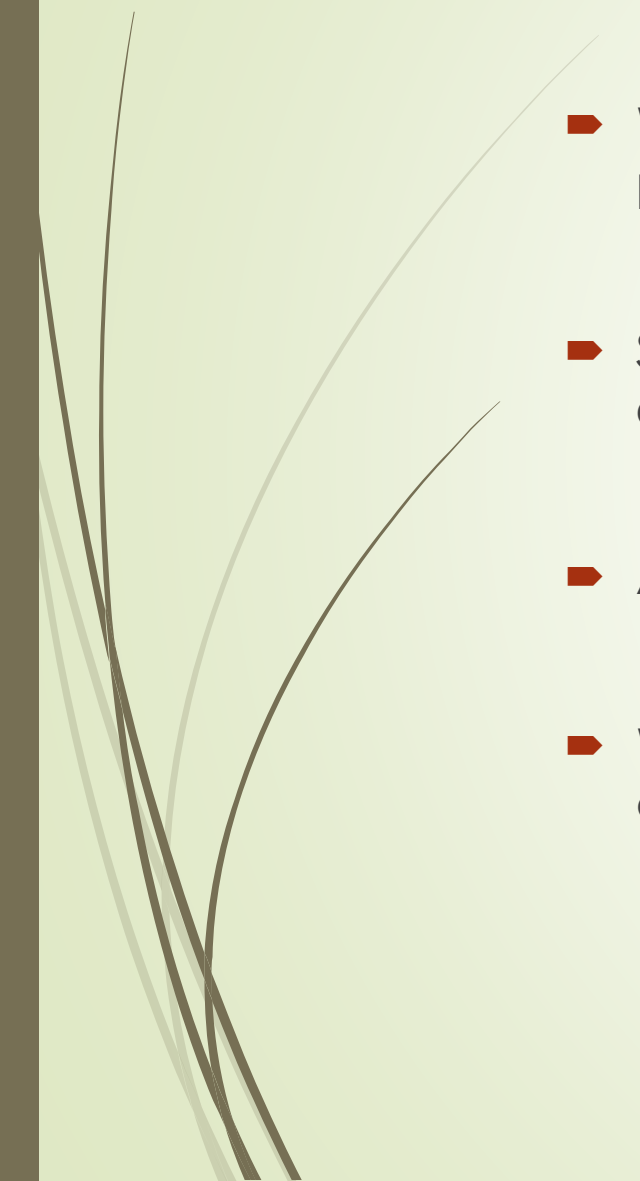


Retrieving Data

- ▶ The dataset that we have used in our project is known as **Credit Card Fraud Detection**.
- ▶ It works on binary classes (1 = Fraud, 0 = Normal) along with 21 other features to identify final label
- ▶ It is basically used to identify frauds in credit card's transaction, whether it is fake or not
- ▶ This dataset largely deals with class imbalance problem as major portion of the data is of label 0 (normal class)

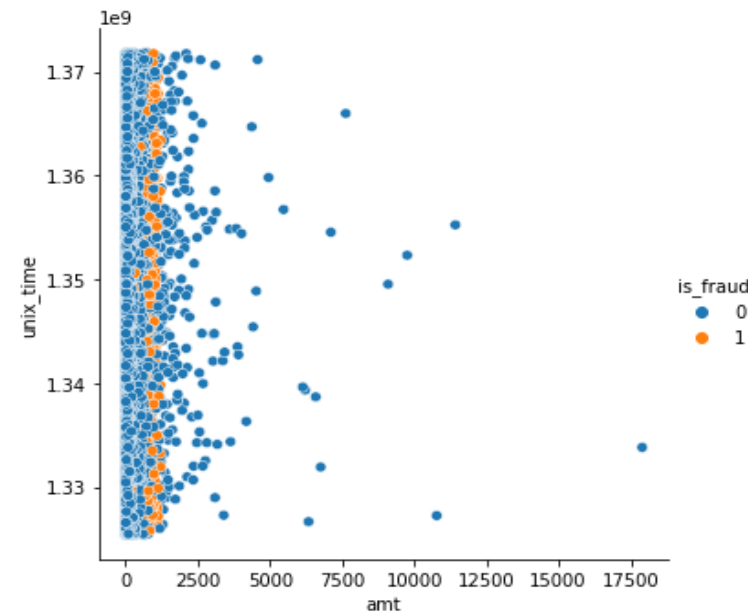
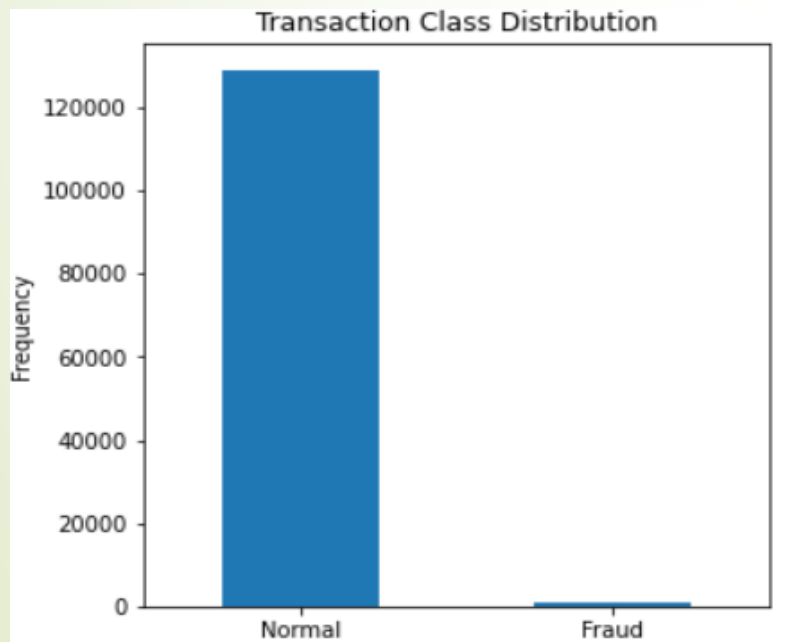


Data Exploration

- ▶ We have total 21 features in our data set bases on which the label is predicted.
 - ▶ Some of the key features are **Latitude, Merchant Name, Transaction amount and Transaction number**
 - ▶ All features contained categorical values
 - ▶ We plotted the heat map to find out the relation of features with each other.
- 

Data Exploration (Cont...)

- We plotted a bar graph to visually represent the imbalance problem as we can see from the figure majority of the data contains **normal** label
- There were **128963** normal cases and only **705** fraud cases in training data





Data Preparation

- In this phase we basically cleaned the data and split the data (train/test)
- We have split the data for training and testing purpose (70% training data and 30% testing data)
- We removed the unwanted columns like **unnamed, first name, last name etc.**
- We converted the Date of Birth to the respective age and replaced the DOB column with age in both training and testing data.
- We created a list of categorical variables and converted them into numerical form by one hot encoding.
- Moreover, we also removed the duplicated columns



Data Modeling

- Initially we used decision tree classifier and logistic regression classifier without balancing the data
- Though we got high accuracy but the data was biased towards normal label so low reliability
- We then performed simple random under sampling on both models which involves randomly selecting examples from the majority class
- After that we performed Inverse Random Under-sampling which works by inverting the cardinalities of majority and minority class.

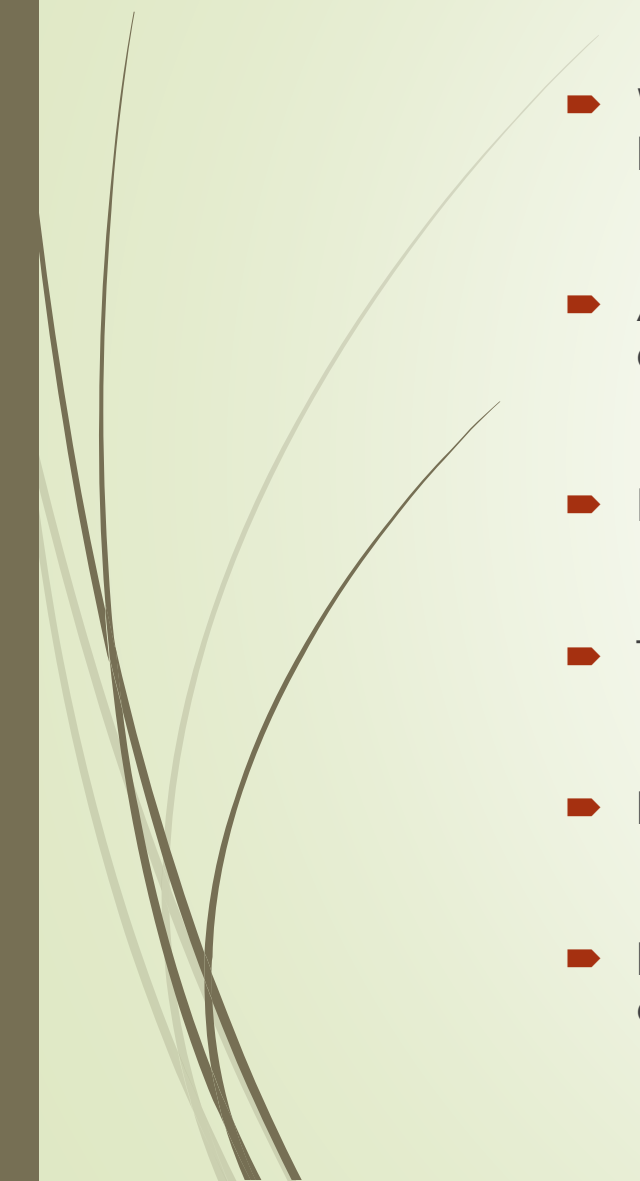


Data Modeling (Cont...)

- The majority class is divided into subsets such that each subset is smaller in size than the minority class
- We have also used bagging technique along with IRUS to increase the true positive rate and decrease the false positive rate
- Through bagging and ensembling technique we get multiple scores depending up on the number of sets and by taking the mean we get the final probability score
- The final probability score is further normalized (z-score) to get same mean and variance and make our classifier more effective
- Furthermore, we set a threshold that if the final score is > 0.5 classify it as **Fraud** else **Normal**



Results

- We employed the F1 score and the ROC-AUC curve to test our model using better techniques of evaluation.
 - As mentioned before we test our results on imbalanced class and then we test our models by using random under sampling techniques.
 - Finally, we put the Inverse Random Model to the test
 - The IRUS model performed well as compared to the other models
 - Inverse Random Under Sampling provides a huge increase in performance
 - It helped to solve the class imbalance problem in much better way as compared to other models.
- 

Results (Cont...)

Models	F1 Score	ROC AUC
Logistic Regression	0.05	0.53
Decision Tree	0.31	0.92
Logistic Regression Under Sampling	0.05	0.57
Decision Tree Under Sampling	0.11	0.95
IRUS	0.46	0.92

Comparative Chart of all models

Model	Recall class-1	Recall class-0	f1-score class-1	f1-score class-0
Logistic Regression	0.00	1.00	0.00	1.00
Decision Tree	0.64	0.99	0.31	0.99
Logistic Regression Under Sample	0.08	0.87	0.01	0.93
Decision Tree Under Sample	0.98	0.92	0.10	0.96

Comparative Chart

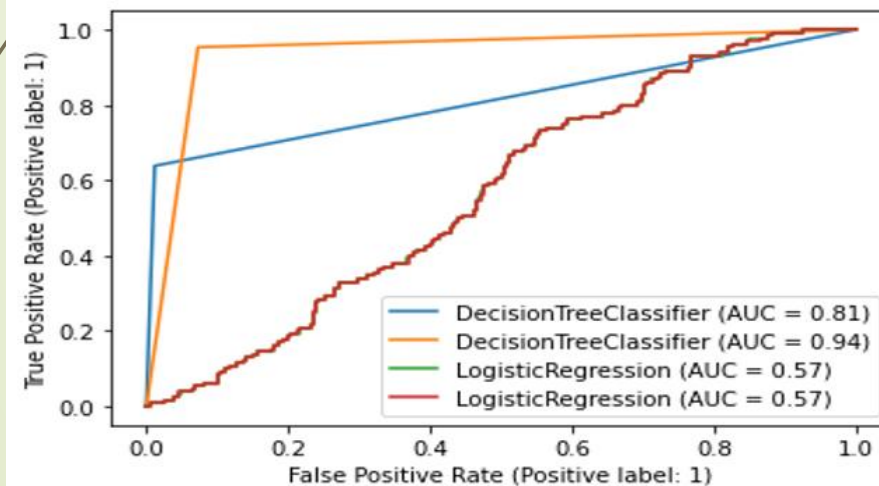


Figure 4: ROC Curve Without IRUS

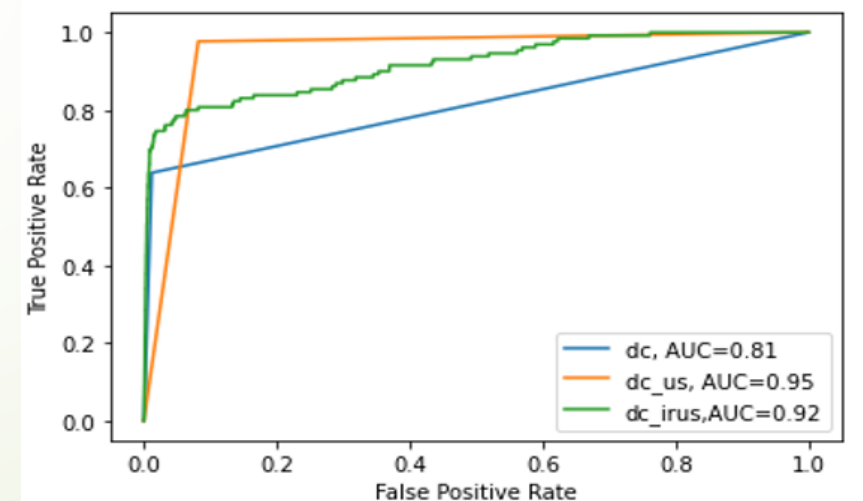


Figure 4: ROC Curve With IRUS