Credit Card Fraud Detection

Team 5

Meet Our Team

Roozbeh Jafari

Yuzhe Zheng

Amber Wu

Phoenix Wang

Jeffrey Leung

Problem Statement

With this dataset, we are hoping to build a model that will successfully predict whether a transaction is fraud or not.

Why should you care?

A growing problem
Credit card fraud increased by 18.4% in 2018 and continues to climb to this day.

Identity theft
This type of fraud accounted for 35.4% of identity theft fraud in 2018.

Money lost
$24.26 Billion was lost worldwide due to credit card fraud in 2018.

Transition to internet
Credit card fraud is here to stay as it is now moving to the digital space.

Affects lives
Criminals can ruin credit scores, making to harder for victims to get loans.

Transaction Processing

What data can we expect to be reviewed?

Exploratory Data Analysis

What can we learn about our dataset?

Highly imbalanced

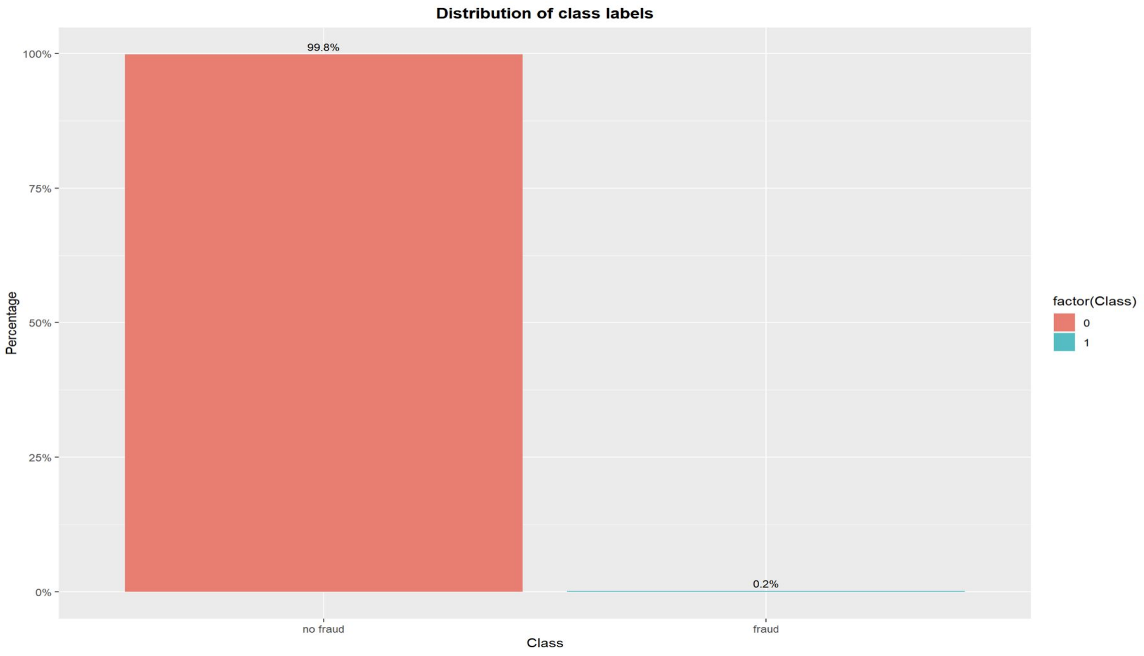
284,315 Non-fraud to **492** fraud.

Time

Non-fraud transactions have a distinct bimo
for time of the transaction.

Fraud Amount

Fraud charges are highly skewed to the left;
were worth 0 dollars.



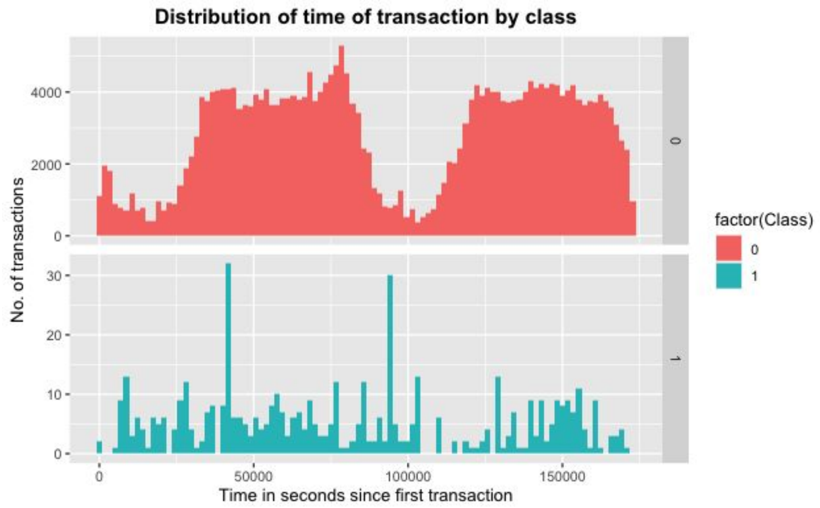
Exploratory Data Analysis

What can we learn about our dataset?

Highly imbalanced
284,315 to 492.

Time
Non-fraud transactions have a distinct bimodal distribution for time of the transaction.

Fraud Amount
Fraud charges are highly skewed to the left; most charges were worth 0 dollars.



Exploratory Data Analysis

What can we learn about our dataset?

Highly imbalanced

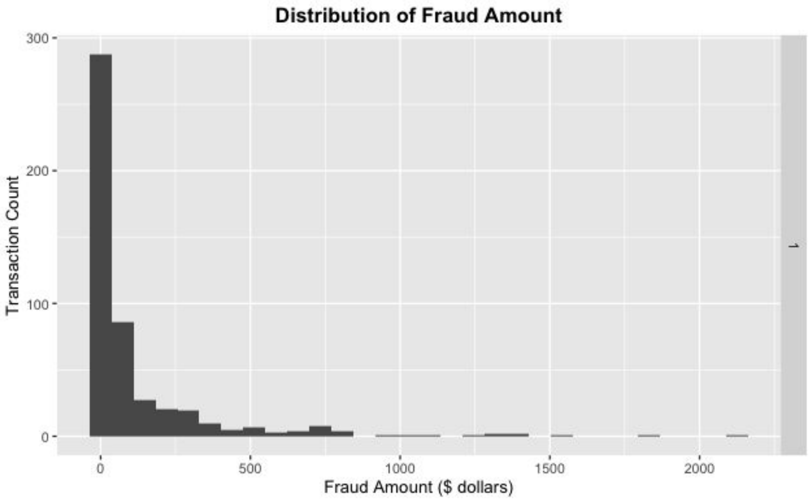
284,315 to 492.

Time

Non-fraud transactions have a distinct bimodal distribution for time of the transaction.

Fraud Amount

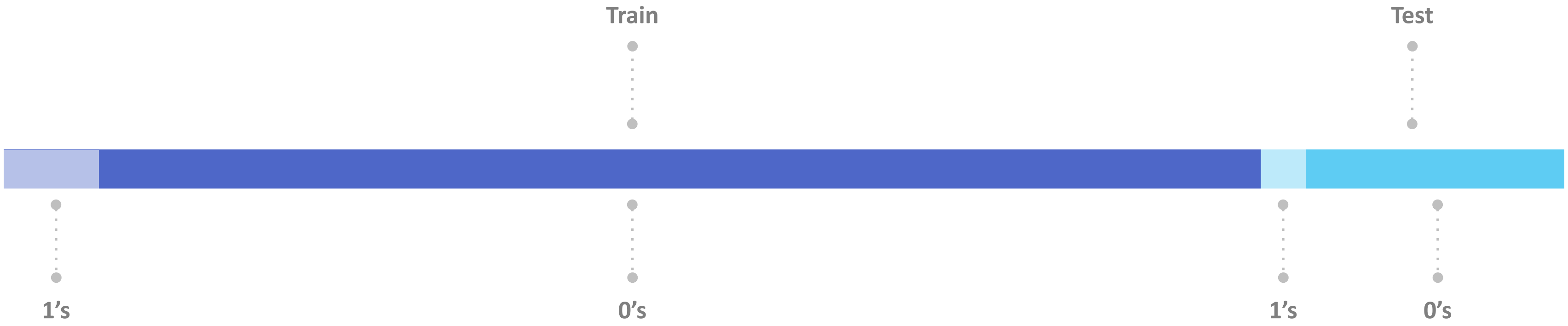
Fraud charges are highly skewed to the left; most charges were worth 0 dollars.



Splitting the Data

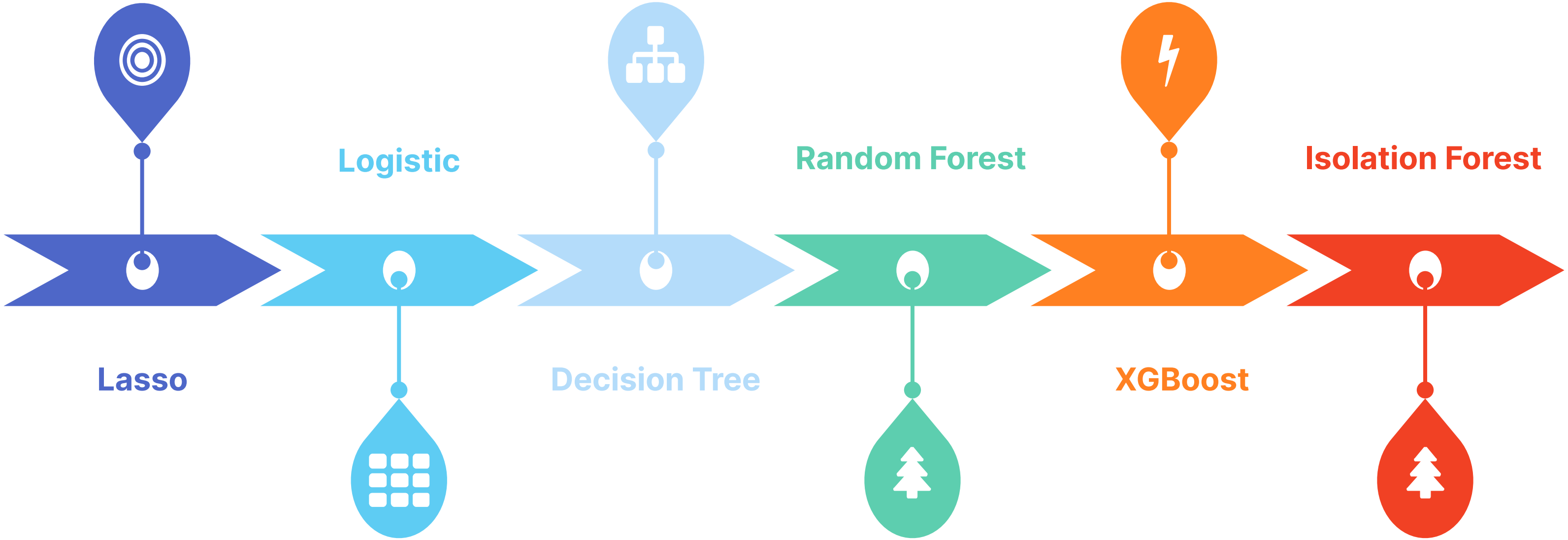
Downsampling to balance our set

Imbalanced



Balanced

Algorithms Used



Lasso Regression

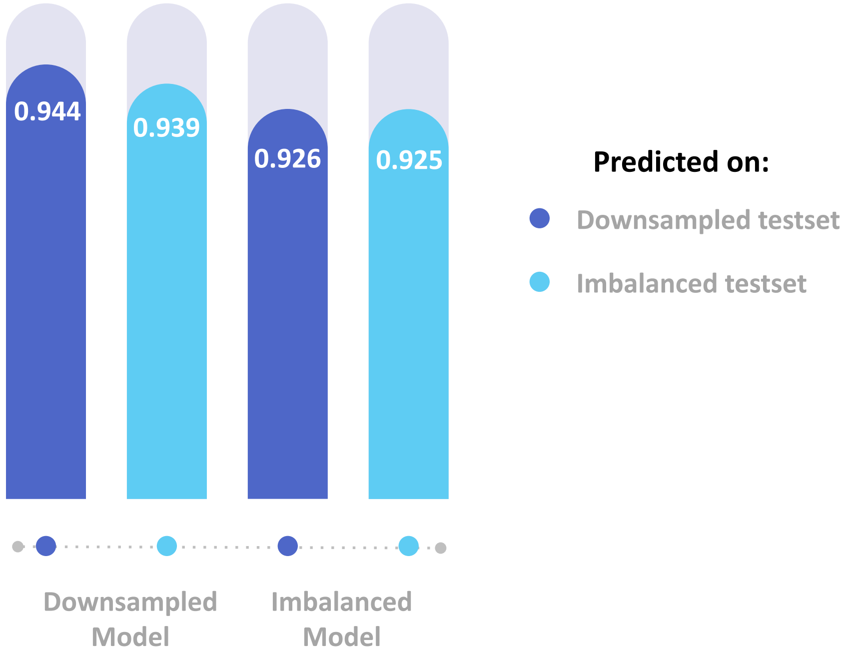
Supervised Learning

Lasso Regression

Area Under the Curve

Area Under Curve

0.944 Out of the 4 models tested, the best performing one was trained on a downsampled training set and tested on the imbalanced testset.



Lasso Regression

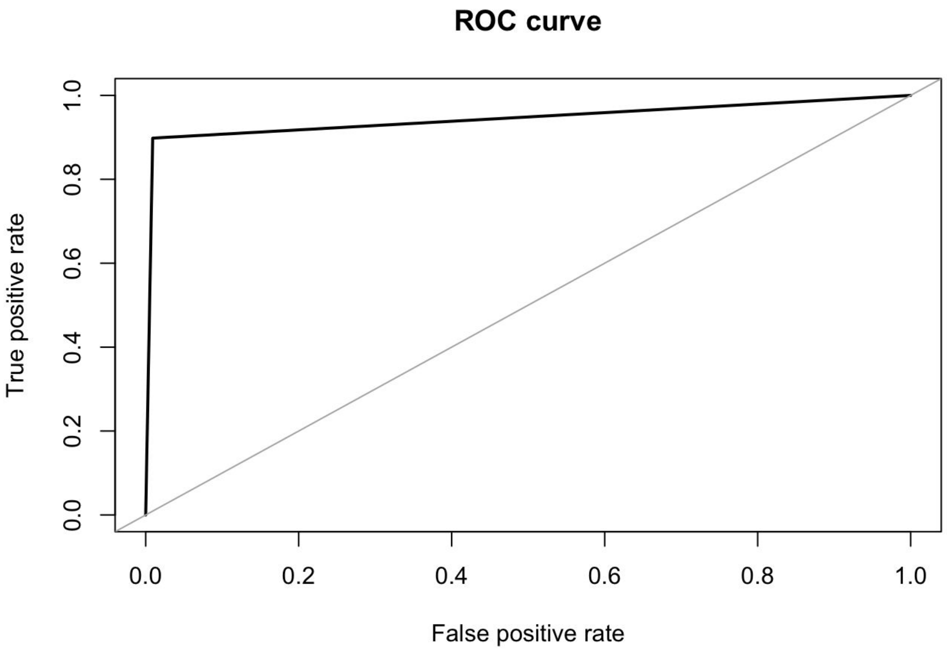
Confusion Matrix - Best Sensitivity

Downsampled Model, Downsampled Data

	Not Fraud	Fraud
Predicted Negative	101	11
Predicted Positive	1	97

89.8% **Sensitivity**
Strong fraud detection

99.1% **Specificity**
Very strong real transaction detection



Logistic Regression

Supervised Learning

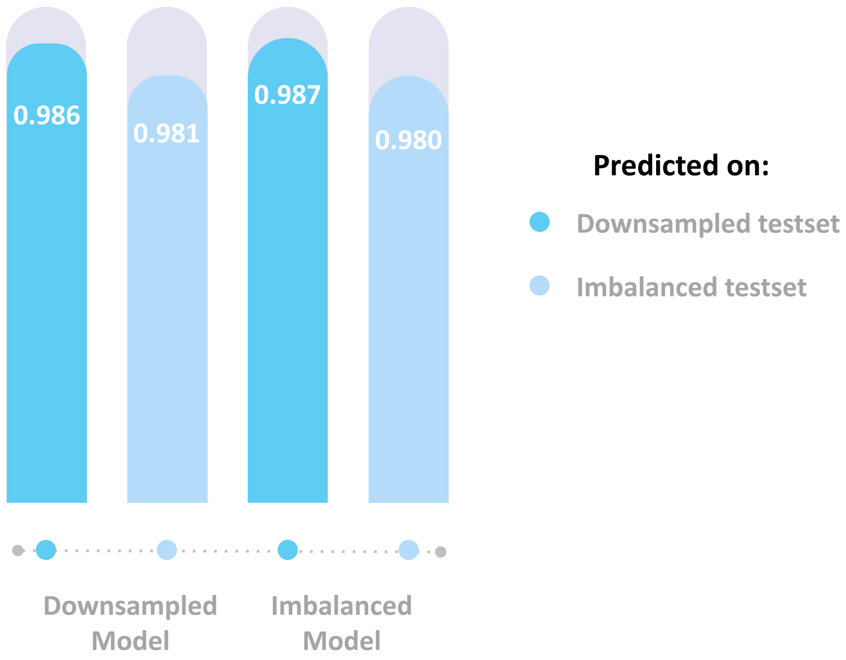
Logistic Regression

Area Under the Curve

Area Under Curve

0.987

Out of the 4 models tested, the best performing one was trained on a downsampled training set and tested on the downsample testset.



Logistic Regression

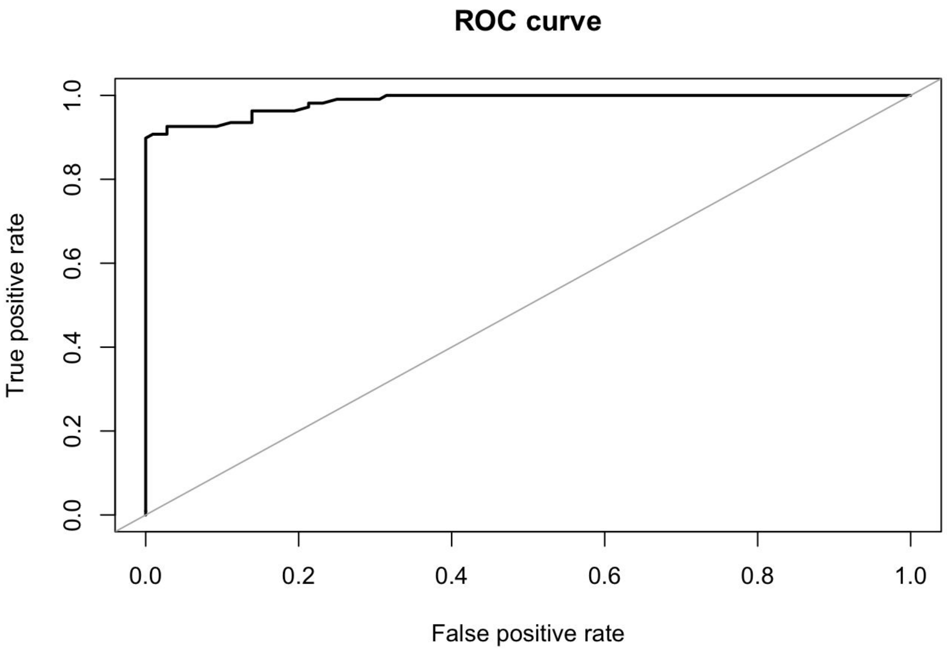
Confusion Matrix - Best Sensitivity

Downsampled Model, Downsampled Data

	Not Fraud	Fraud
Predicted Negative	105	10
Predicted Positive	3	98

90.7% **Sensitivity**
Strong fraud detection

97.2% **Specificity**
Very strong real transaction detection



Decision Tree

Supervised Learning

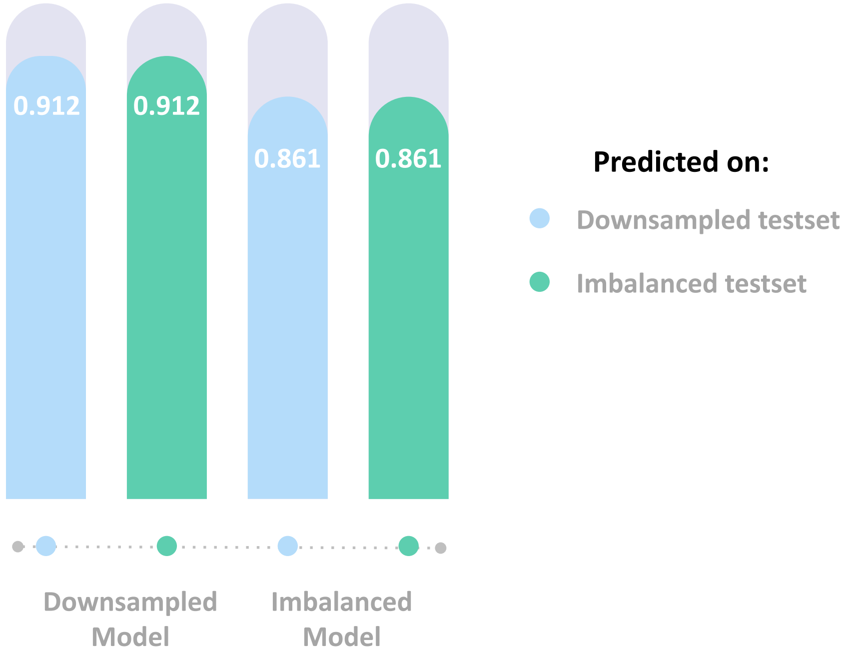
Decision Tree

Area Under the Curve

Area Under Curve

0.912

Out of the 4 models tested, the best performing one was trained on a downsampled training set and tested on the imbalanced testset.



Decision Tree

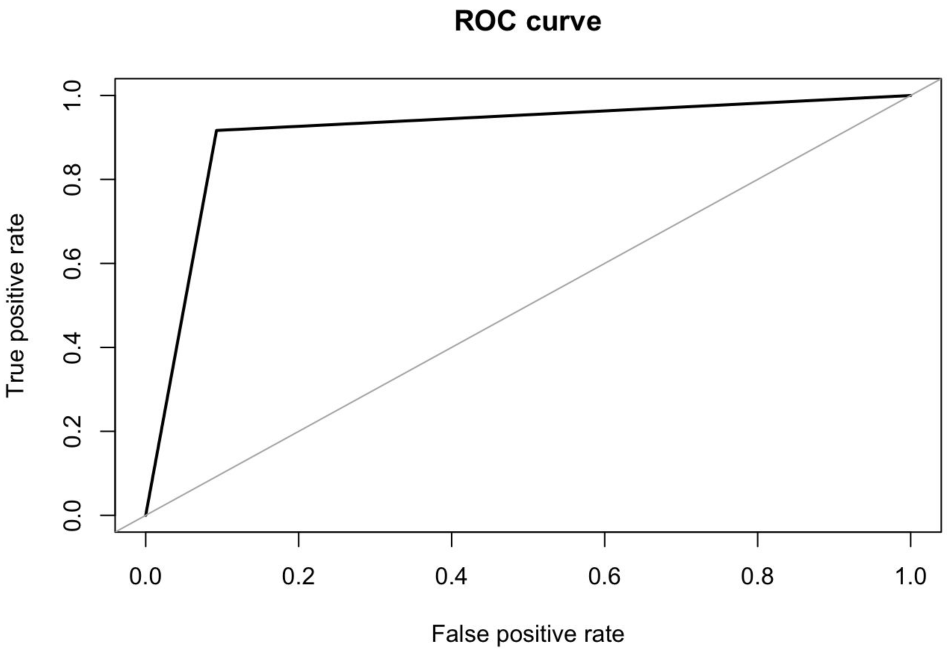
Confusion Matrix - Best Sensitivity

Downsampled Model, Downsampled Data

	Not Fraud	Fraud
Predicted Negative	98	9
Predicted Positive	10	99

91.7% **Sensitivity**
Strong fraud detection

90.7% **Specificity**
Strong real transaction detection



Random Forest

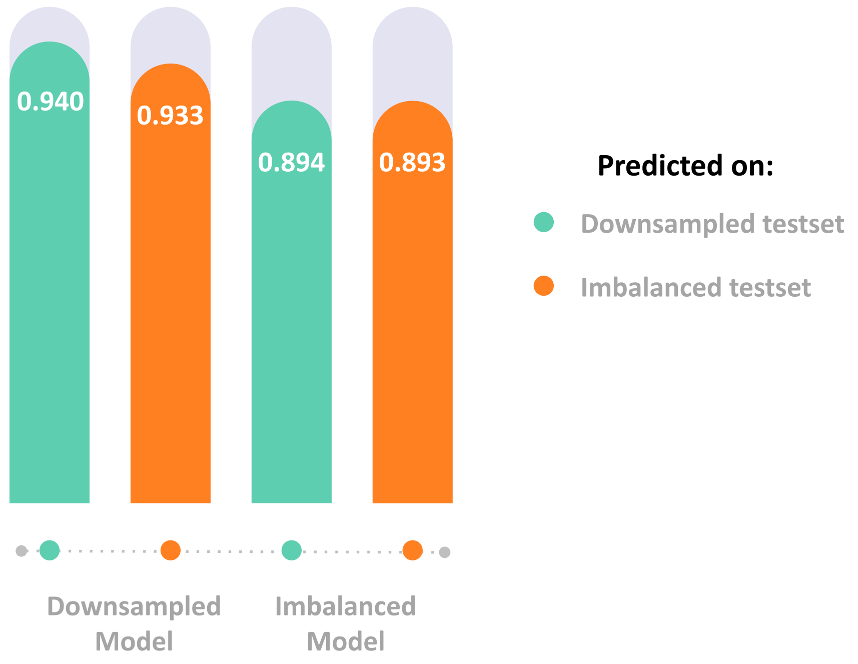
Supervised Learning

Random Forest

Area Under the Curve

Area Under Curve

0.940 Out of the 4 models tested, the best performing one was trained on a downsampled training set and tested on the imbalanced testset.



Random Forest

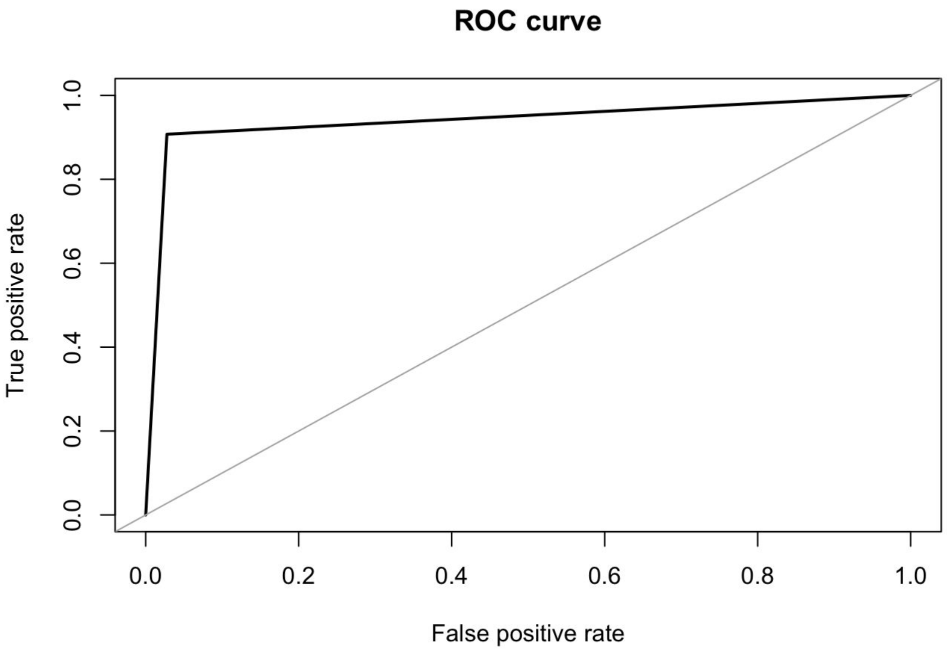
Confusion Matrix - Best Sensitivity

Downsampled Model, Downsampled Data

	Not Fraud	Fraud
Predicted Negative	105	10
Predicted Positive	3	98

90.7% **Sensitivity**
Strong fraud detection

97.2% **Specificity**
Very strong real transaction detection



XGBoost

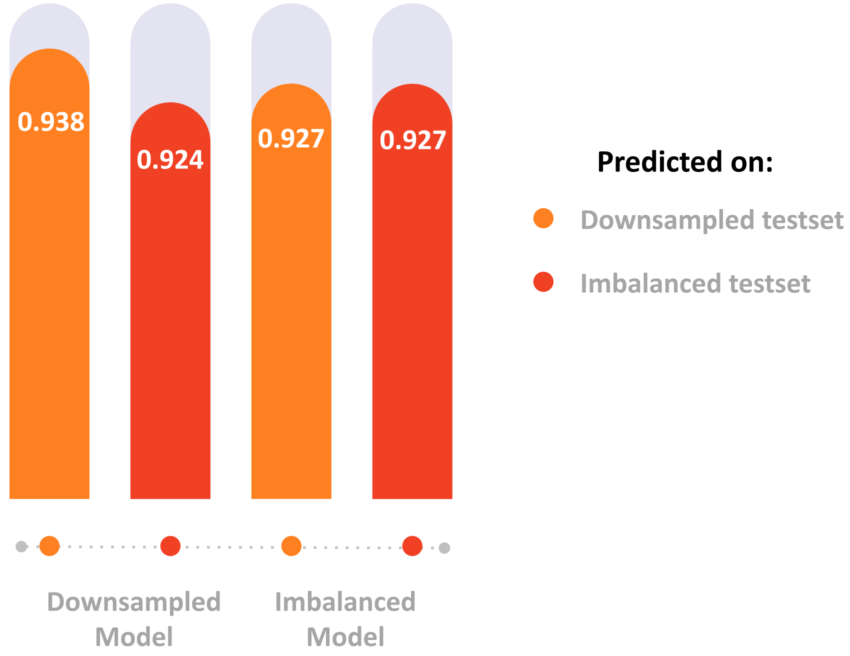
Supervised Learning

XGBoost

Area Under the Curve

Area Under Curve

0.938 Out of the 4 models tested, the best performing one was trained on a downsampled training set and tested on the downsampled testset.



XGBoost

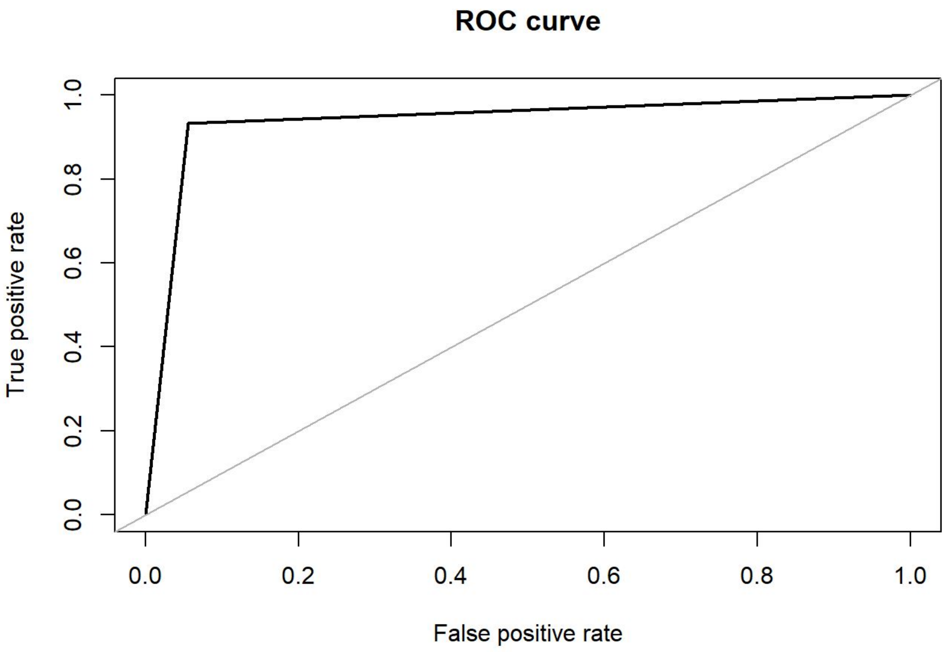
Confusion Matrix - Best Sensitivity

Downsampled Model, Downsampled Data

	Not Fraud	Fraud
Predicted Negative	84	6
Predicted Positive	5	83

93.3% **Sensitivity**
Strong fraud detection

94.4% **Specificity**
Strong real transaction detection



Isolation Forest

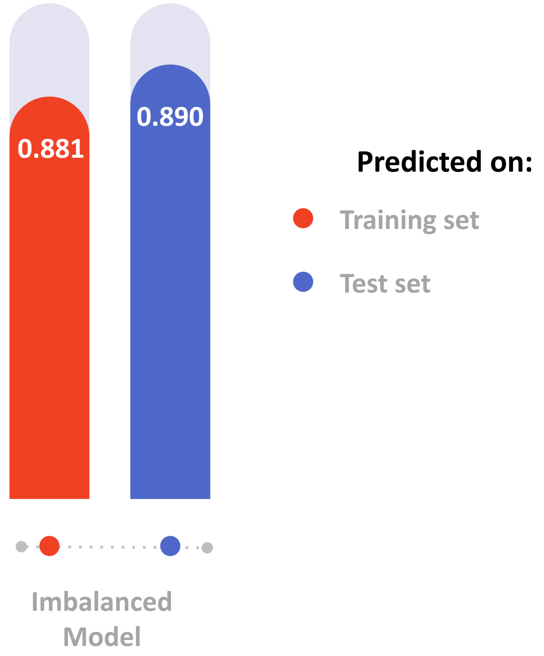
Unsupervised Learning

Isolation Forest

Area Under the Curve

Area Under Curve

0.890 Out of the 2 models tested, the best performing one was trained on a downsampled training set and tested on the imbalanced testset.



Isolation Forest

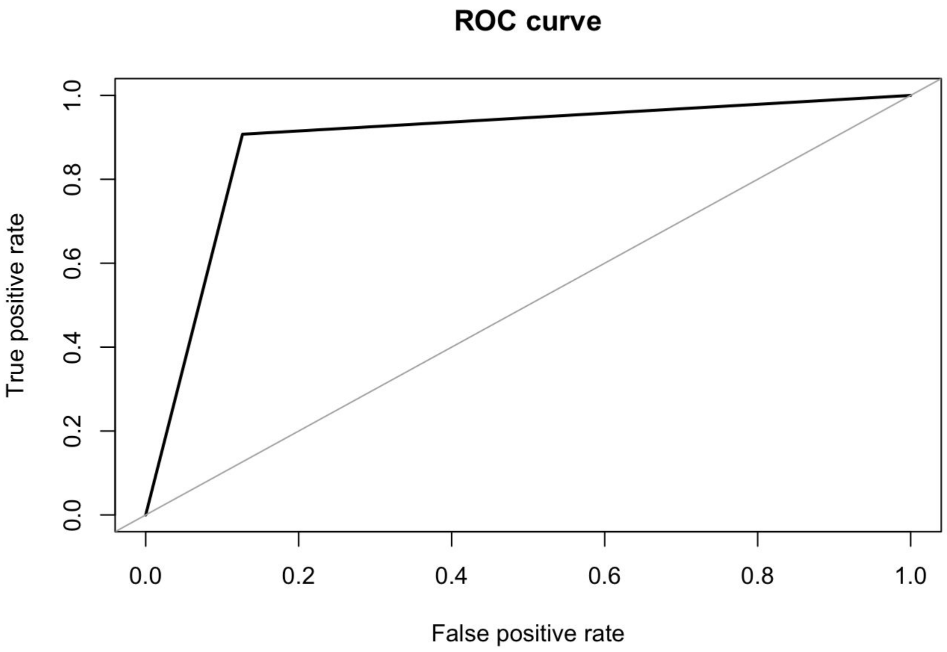
Confusion Matrix - Best Sensitivity

Training Model, Testing Data

	Not Fraud	Fraud
Predicted Negative	49656	10
Predicted Positive	7197	98

90.7% **Sensitivity**
Strong fraud detection

87.3% **Specificity**
Moderate real transaction detection



Challenges

Imbalanced Dataset

Going into this project, we had concerns over getting effective results with our imbalanced dataset, but downsampling proved effective.

Large Dataset

There were instances of algorithms that took too long to run due to the sheer size of the data set. For one of the algorithms, we further reduced the imbalance set.

Nameless Variables

Without known variables, it is hard to interpret our dataset well. Because of this, we chose to make as accurate predictions as possible.

Performance Overview

How did we do overall?

Below is a matrix comparing all of our models. While we ran multiple models for each algorithm, we choose to select the ones that had the highest sensitivity.

	AUC	Sensitivity	Specificity
Lasso Regression	0.944	0.898	0.991
Logistic Regression	0.986	0.907	0.972
Decision Tree	0.912	0.917	0.907
Random Forest	0.940	0.907	0.973
XGBoost	0.938	0.933	0.944
Isolation Forest	0.890	0.907	0.874

Key Learnings

What can we take away from this project?

Models were effective

All of our models had AUC and sensitivity of 0.89 and above.

Not hindered by imbalanced data

By downsampling, we were able to get the algorithms to learn fraud charges as opposed to it learning non-fraud charges.

The sensitivity-specificity tradeoff

When choosing a threshold, we had to consider the cost of false positives and false negatives.

Credit card fraud theft is a serious issue

It is one of the fastest-growing forms of identity theft, which is why it is important that we can effectively utilize machine learning to predict when it happens.

Thank You

Happy Modeling