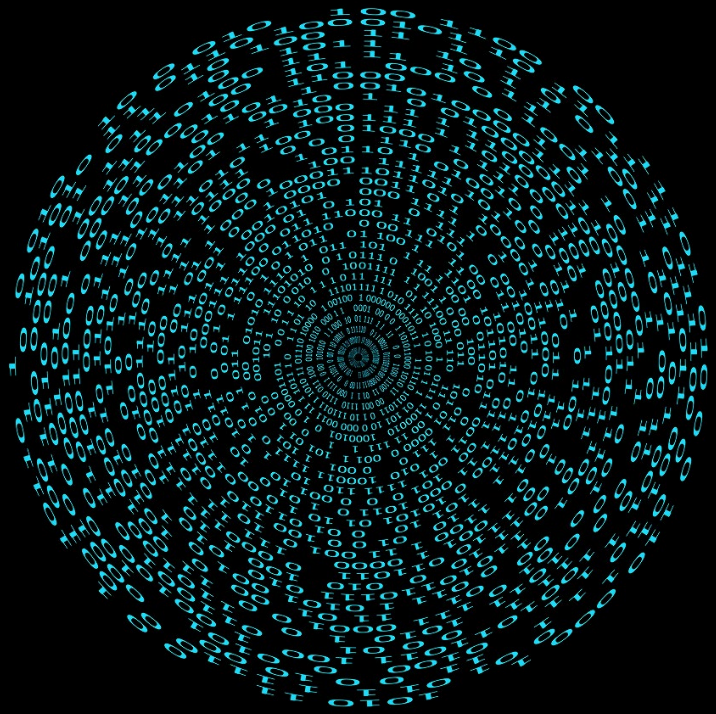
Deloitte.



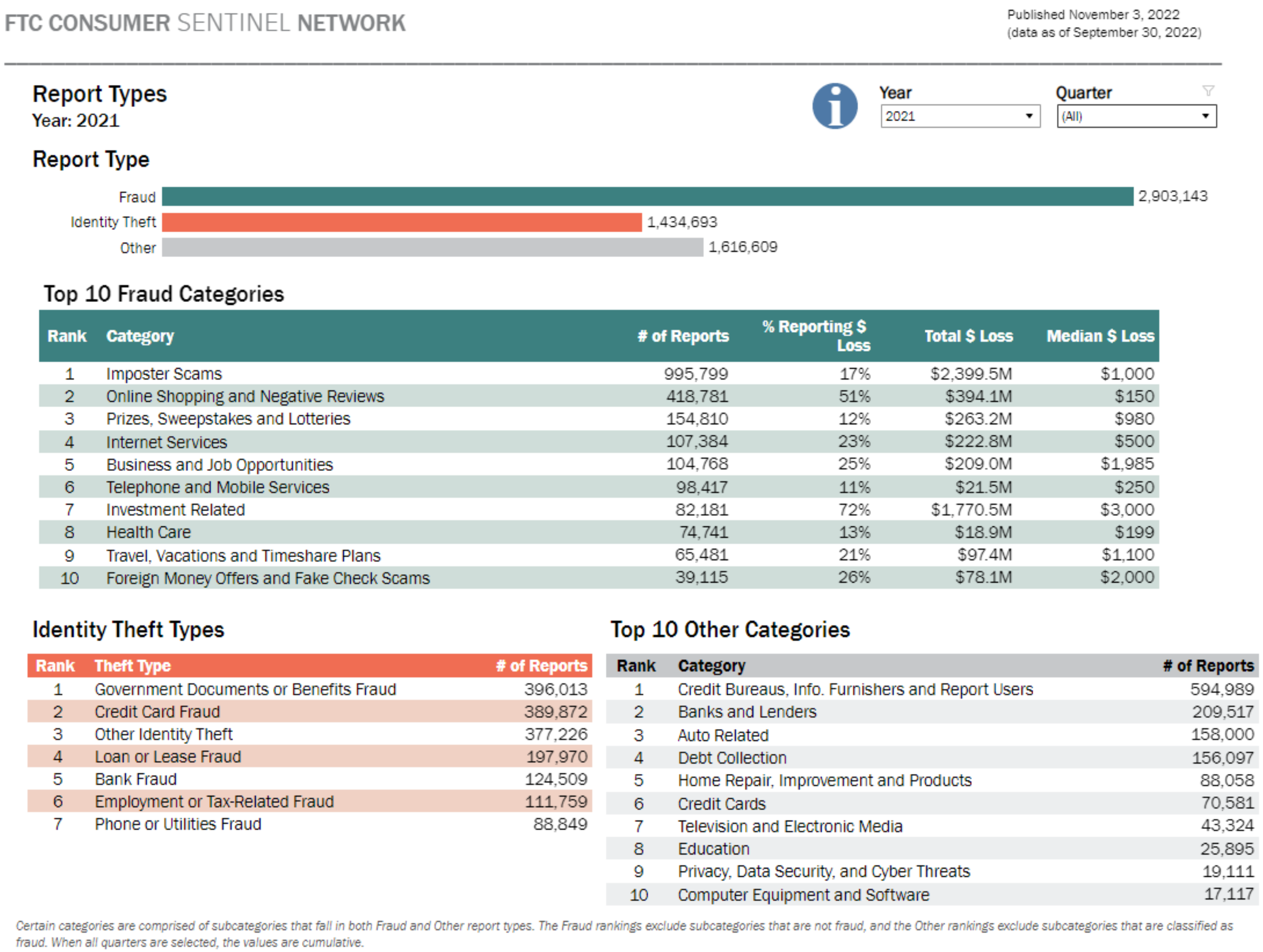
# AI Academy Capstone – Found in a Random Forest

MAKING AN IMPACT THAT MATTERS since 1845

February 3, 2023

# How prevalent is financial fraud?

Organizations, governments, and private citizens are at risk of becoming victims of fraud.



FTC CONSUMER SENTINEL NETWORK

Published November 3, 2022
(data as of September 30, 2022)

## Report Types

Year: 2021

Year: 2021 ▾ Quarter: (All) ▾

### Report Type

| Report Type | |
|---|---|
| Fraud | 2,903,143 |
| Identity Theft | 1,434,693 |
| Other | 1,616,609 |

### Top 10 Fraud Categories

| Rank | Category | # of Reports | % Reporting $ Loss | Total $ Loss | Median $ Loss |
|---|---|---|---|---|---|
| 1 | Imposter Scams | 995,799 | 17% | $2,399.5M | $1,000 |
| 2 | Online Shopping and Negative Reviews | 418,781 | 51% | $394.1M | $150 |
| 3 | Prizes, Sweepstakes and Lotteries | 154,810 | 12% | $263.2M | $980 |
| 4 | Internet Services | 107,384 | 23% | $222.8M | $500 |
| 5 | Business and Job Opportunities | 104,768 | 25% | $209.0M | $1,985 |
| 6 | Telephone and Mobile Services | 98,417 | 11% | $21.5M | $250 |
| 7 | Investment Related | 82,181 | 72% | $1,770.5M | $3,000 |
| 8 | Health Care | 74,741 | 13% | $18.9M | $199 |
| 9 | Travel, Vacations and Timeshare Plans | 65,481 | 21% | $97.4M | $1,100 |
| 10 | Foreign Money Offers and Fake Check Scams | 39,115 | 26% | $78.1M | $2,000 |

### Identity Theft Types

| Rank | Theft Type | # of Reports |
|---|---|---|
| 1 | Government Documents or Benefits Fraud | 396,013 |
| 2 | Credit Card Fraud | 389,872 |
| 3 | Other Identity Theft | 377,226 |
| 4 | Loan or Lease Fraud | 197,970 |
| 5 | Bank Fraud | 124,509 |
| 6 | Employment or Tax-Related Fraud | 111,759 |
| 7 | Phone or Utilities Fraud | 88,849 |

### Top 10 Other Categories

| Rank | Category | # of Reports |
|---|---|---|
| 1 | Credit Bureaus, Info. Furnishers and Report Users | 594,989 |
| 2 | Banks and Lenders | 209,517 |
| 3 | Auto Related | 158,000 |
| 4 | Debt Collection | 156,097 |
| 5 | Home Repair, Improvement and Products | 88,058 |
| 6 | Credit Cards | 70,581 |
| 7 | Television and Electronic Media | 43,324 |
| 8 | Education | 25,895 |
| 9 | Privacy, Data Security, and Cyber Threats | 19,111 |
| 10 | Computer Equipment and Software | 17,117 |

*Certain categories are comprised of subcategories that fall in both Fraud and Other report types. The Fraud rankings exclude subcategories that are not fraud, and the Other rankings exclude subcategories that are classified as fraud. When all quarters are selected, the values are cumulative.*

1. More than **$5.8 billion** was lost due to report fraud.
2. ~ **390,000 reports** of credit card fraud.
3. Financial institutions aren’t the only organizations impacted.

Source: FTC Consumer Sentinel Network, 2022

# What data set was used and what models were utilized?

The decision tree and random forest analytical module were trained on a real-world credit card transaction data set.

## Credit Card Fraud Detection (Kaggle)

- The dataset used contains transactions made by credit cards in September 2013 by European cardholders.
- The dataset presents 284,807 transactions, of which 492 were fraudulent.

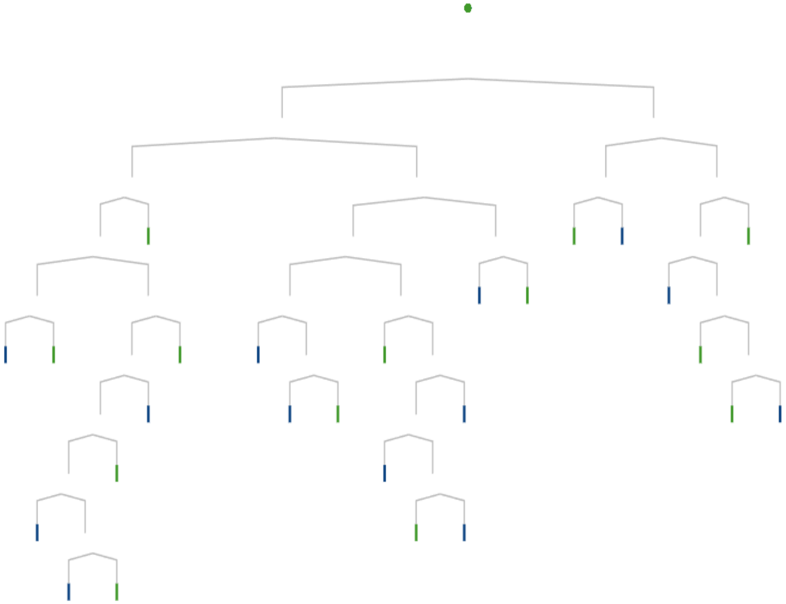
## Features of the Data Set

- 28 of 30 features have already undergone a PCA transformation.
- Due to confidentiality reasons, the only labeled features are 'Time' and 'Amount', which are the two that have not undergone PCA.
- The 'Class' feature is the response variable.

## Modules

## Decision Tree Analytical Module

- Supervised machine learning algorithm.
- A decision tree is a flowchart-like structure in which each internal node represents a test on a feature.
- Each leaf node represents a class label.
- Great for classification problems and predictive analysis as they break down complex data into more manageable information.
- They are easily interpretated.

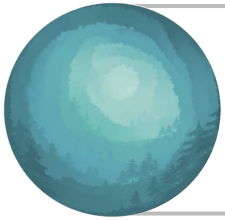


## Random Forest Analytical Module

- Supervised machine learning algorithm.
- A random forest consists of a number of individual decision trees that operate as an ensemble.
- Each individual tree in the forest performs its own class prediction.
- A random forest is the natural progression from decision trees as they reduce the risk of overfitting and can easily determine and evaluate feature importance.

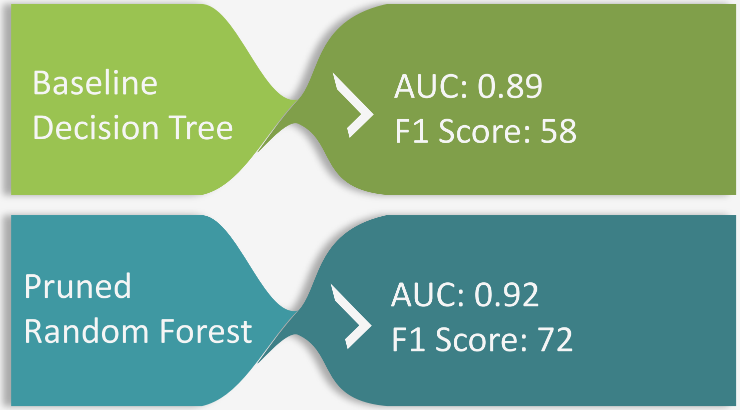
# Impact of the Models | Results Overview

The pruned random forest model performed better than the baseline decision tree and was less complex.



Accuracy: The number of true predictions over the number of predictions.
Area Under the Curve (AUC): The aggregate measure of performance across all possible classification thresholds.
F1 Score: The harmonic mean of precision and recall.

## Results



Baseline Decision Tree

AUC: 0.89
F1 Score: 58

Pruned Random Forest

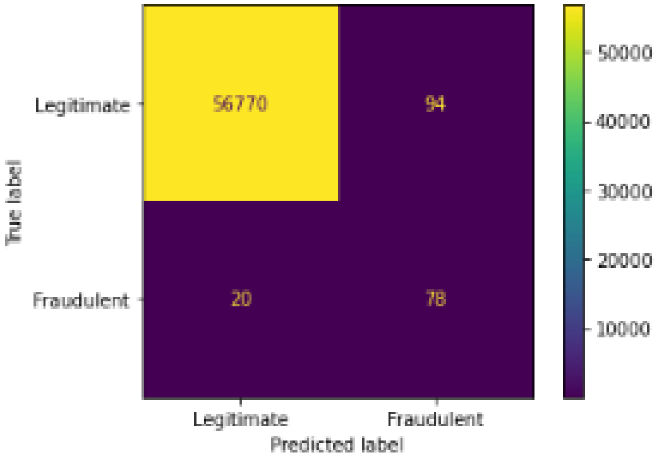
AUC: 0.92
F1 Score: 72

### Higher Performance and Less Complexity

The pruned random forest model only utilized the **top six performing features** from the vanilla random forest model, while the baseline decision tree utilized all 30 features.

# Baseline Decision Tree Model

The model was most likely overfitting and had a false positive rate of 0.00165 and a false negative rate of 3.5e-4.



Accuracy
99.8
%

Precision
47

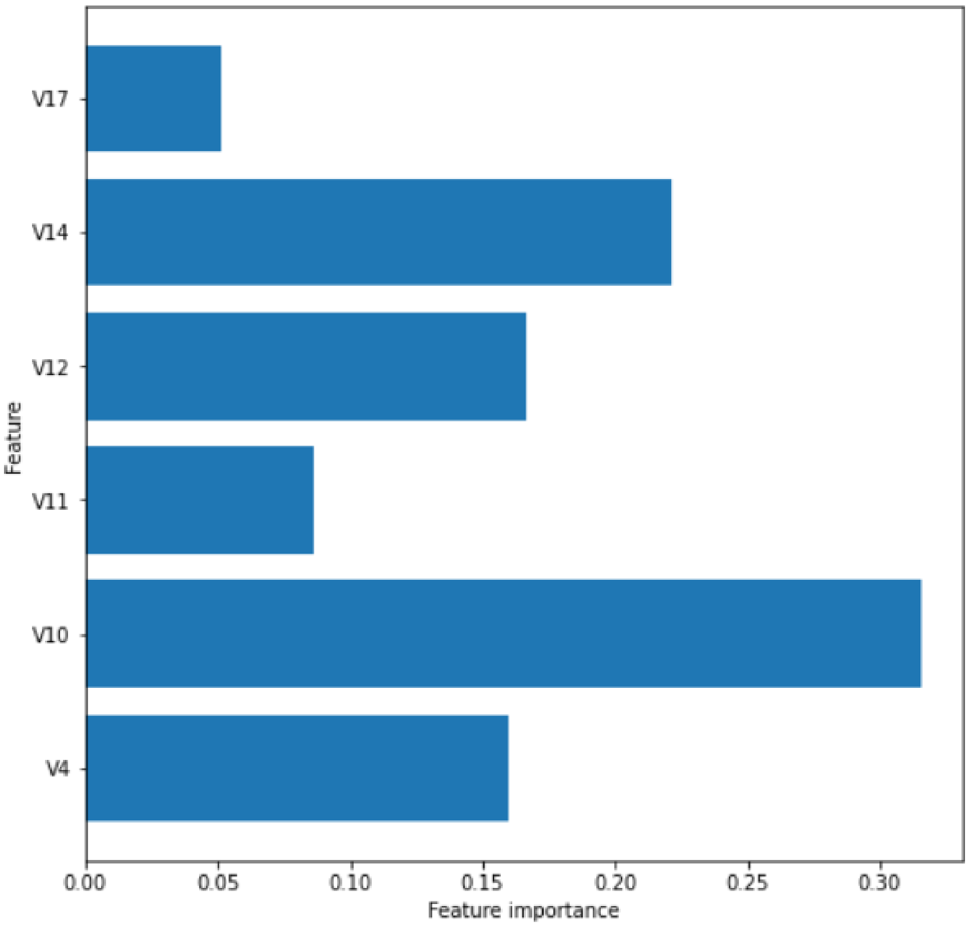
Recall
78

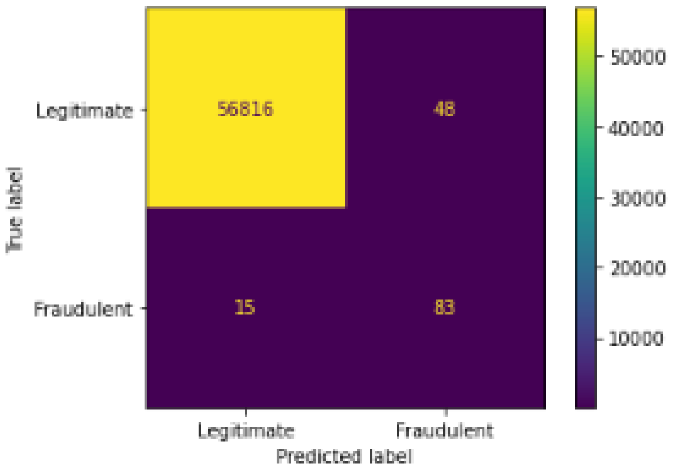
F1 Score
58

AUC
0.89

# Pruned Random Forest Model

The model was less complex and had a false positive rate of 8.4e-4 and false negative rate of 2.6e-4.





Accuracy

99.9 %

Precision

63

Recall

84

F1 Score

72

AUC

0.92

# Impact of the Models | Conclusion

The pruned random forest model performed better than the baseline decision tree and was less complex.

### Higher Performance

- Accuracy: 99.9%
- Precision: 63
- Recall: 84
- F1 Score: 72
- AUC: 0.92

### Reduced Complexity

- The model was reduced from 30 features to just the six top performing features.
- With less features, the model ran faster and was easier to evaluate.

### Low False Negative Rate

- When wrong, the model tends to lean toward false positive over false negative, which in the case of fraud is the more desirable outcome.

### Recommendation

It is recommended that the model be used in a **cursory fashion** to quickly flag transactions for further review. The role of the model is not to serve as the jury, judge, and executioner, but rather, serve as a tool for investigators to more efficiently review transactions and be the first line of defense against fraud.

# What more can be done to improve the model and combat financial fraud?

XGBoosting the random forest model could improve performance and neural networks can be utilized.

### XGBoost

Extreme Gradient Boosting

Good for large data sets

Residual trees are built by calculating similarity scores

### Neural Network

Utilizes interconnected nodes and hidden layers

Good for complex and large data sets

Deloitte.

# Thank you

**Questions…**