Credit Card Fraud Detection

Presented by

Sadvi Sandhya

Agenda

- Business Impact and Problem Definition
- Approach, Design, Methodology
- Value Delivery
- Summary and Conclusions

Presented by
Sadvi Sandhya

Business Impact

By 2027, financial service providers are expected to take a $40 billion hit in credit card losses, a significant increase compared to $28 billion in 2018.

The coronavirus pandemic is also fueling explosive growth in card fraud activity.

With global credit card fraud loss on the rise, it is important for banks, as well as e-commerce companies, to be able to detect fraudulent transactions.

Presented by
Sadvi Sandhya

Business Questions

- With the cost of fraud rising and cardholder trust declining, financial institutions need to take steps to ensure their business and their cardholders are protected.

- How to detect the fraud transactions ?

- How Machine learning models can effectively detect the fraud transactions ?

- Are they better than the rule based fraud detection system?

- Stake holder wants to reduce their annual loss due to credit card fraud which is GBP 5.5M at present

Business Domain : Credit card Union / Bank

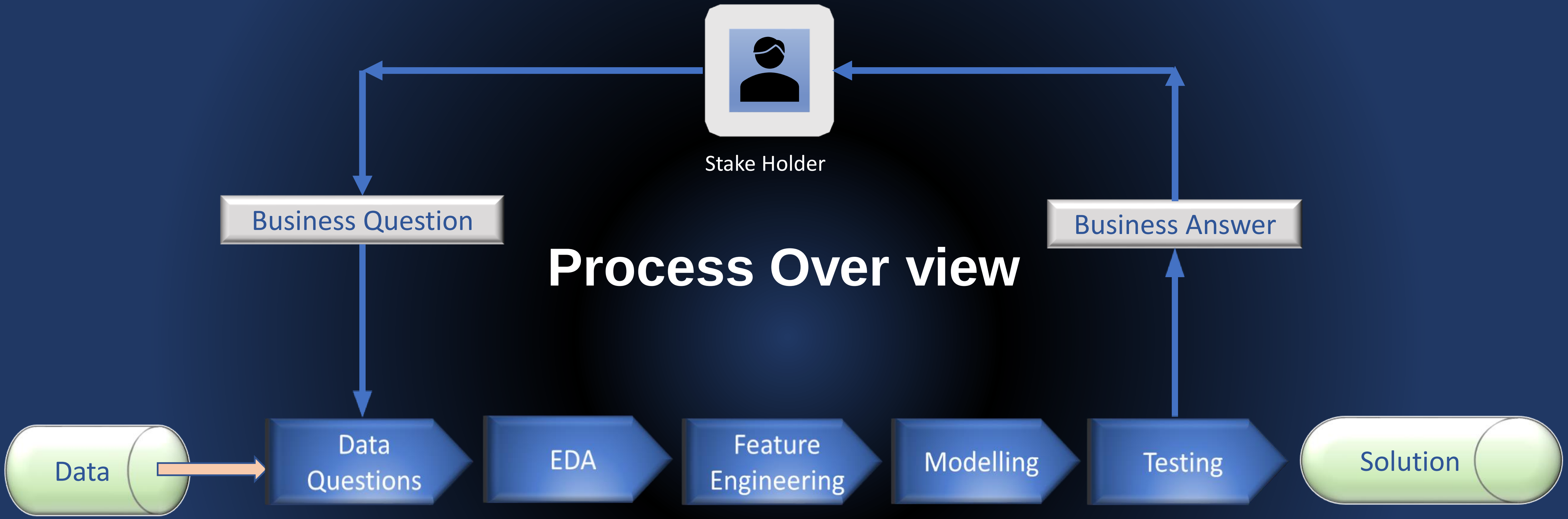
Stake holder : Senior Vice president / Director Risk Management

Presented by

Sadvi Sandhya

Design

Process Over view



Stake Holder

Business Question

Business Answer

Data

Data Questions

EDA

Feature Engineering

Modelling

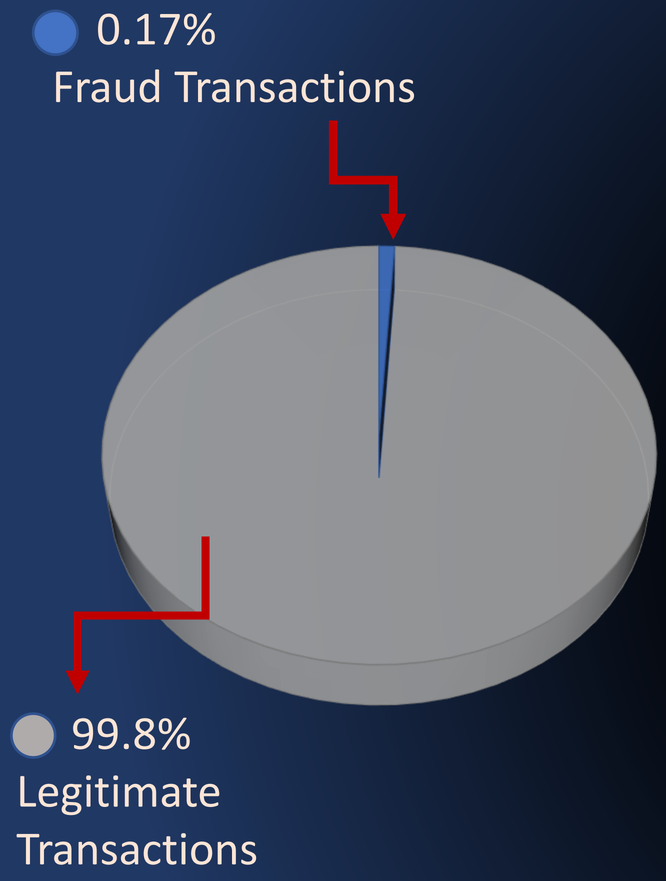
Testing

Solution

Presented by

Sadvi Sandhya

The Dataset



- The dataset is provided by Kaggle and contains transactions made by credit cards in September 2013 by European cardholders.
- Dataset presents transactions that occurred in two days
- Heavily imbalanced , labelled dataset with 492 fraud transactions out of 284,807 transactions.
- Clean data no null values
- There are 31 features out of which 28 are PCA transformed
- Feature 'Class' is the target variable and it takes value 1 in case of fraud and 0 otherwise. Hence it is a Binary Classification Task
- 70% Training the model and 30% testing

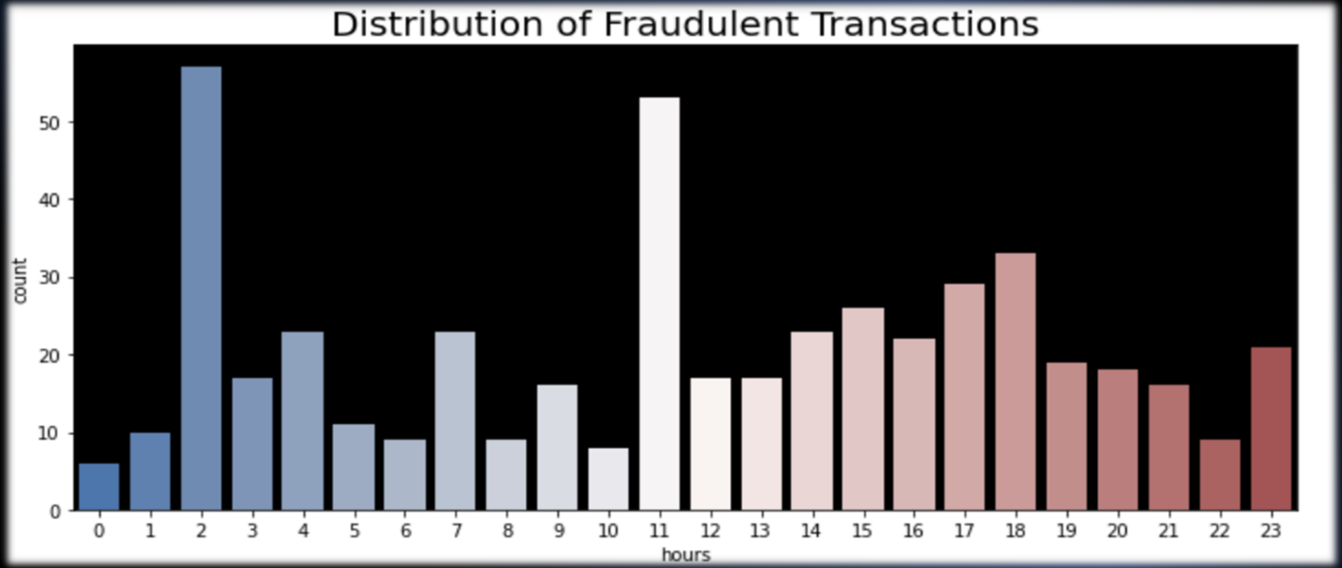
Presented by

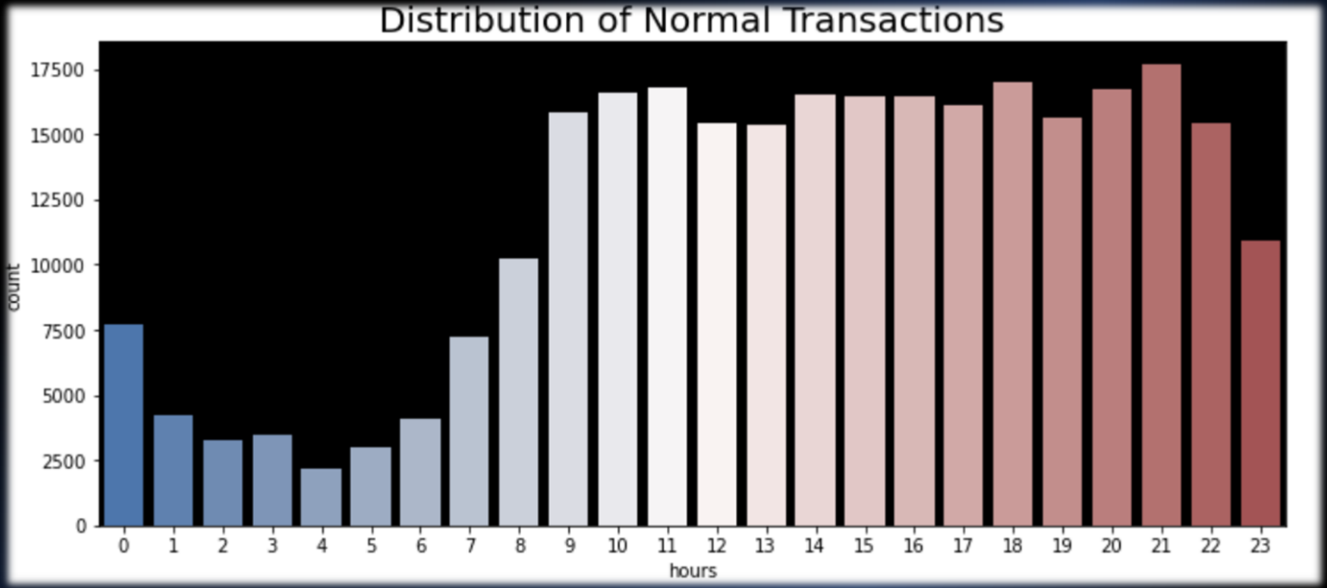
Sadvi Sandhya

Data Analysis

Observations on time of transactions

- Dataset presents transactions that occurred in two days
- Less number of Non fraud transactions from hour 1 to 6 where as there are considerable number of Fraud transactions from hour 1 to 6





Presented by
Sadvi Sandhya

Distribution of Time Vs Transaction Amount

Presented by
Sadvi Sandhya

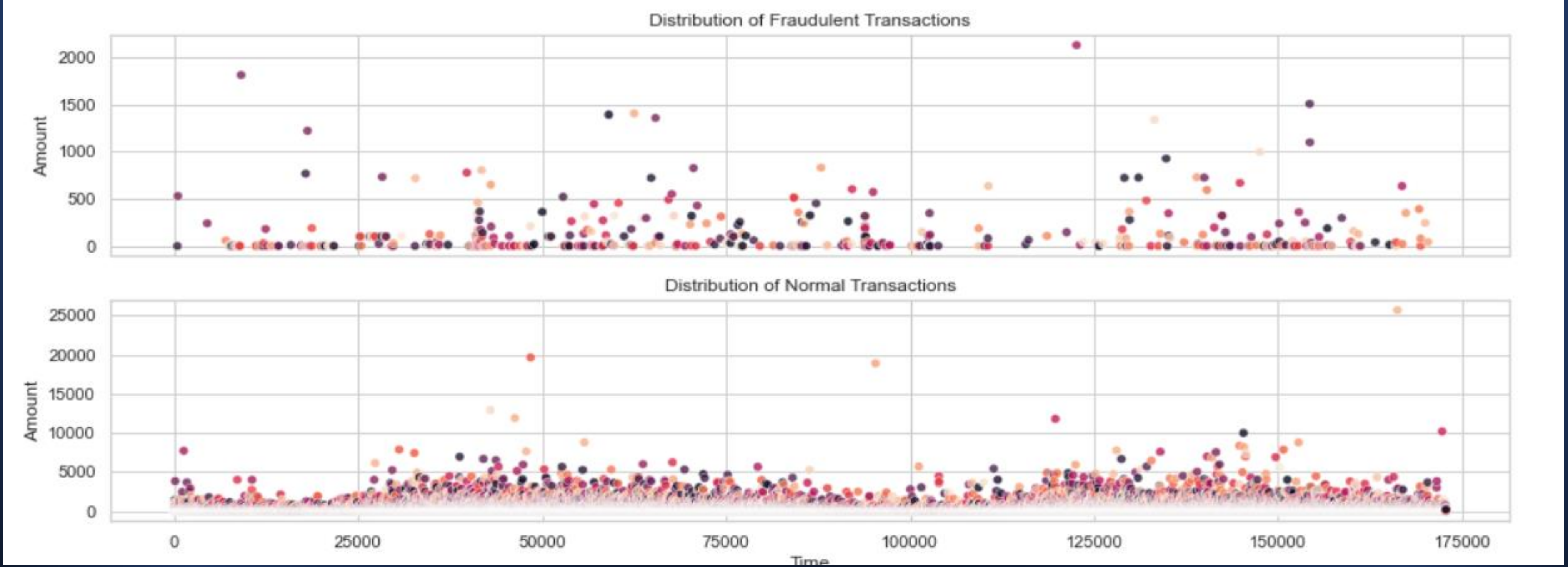
FRAUDULENT

- ► There are much more outliers as compared to normal transactions.

- ► The plot seems to not have any inherent pattern.

NORMAL

- ► There are a less number of outliers as compared to fraudulent transactions.

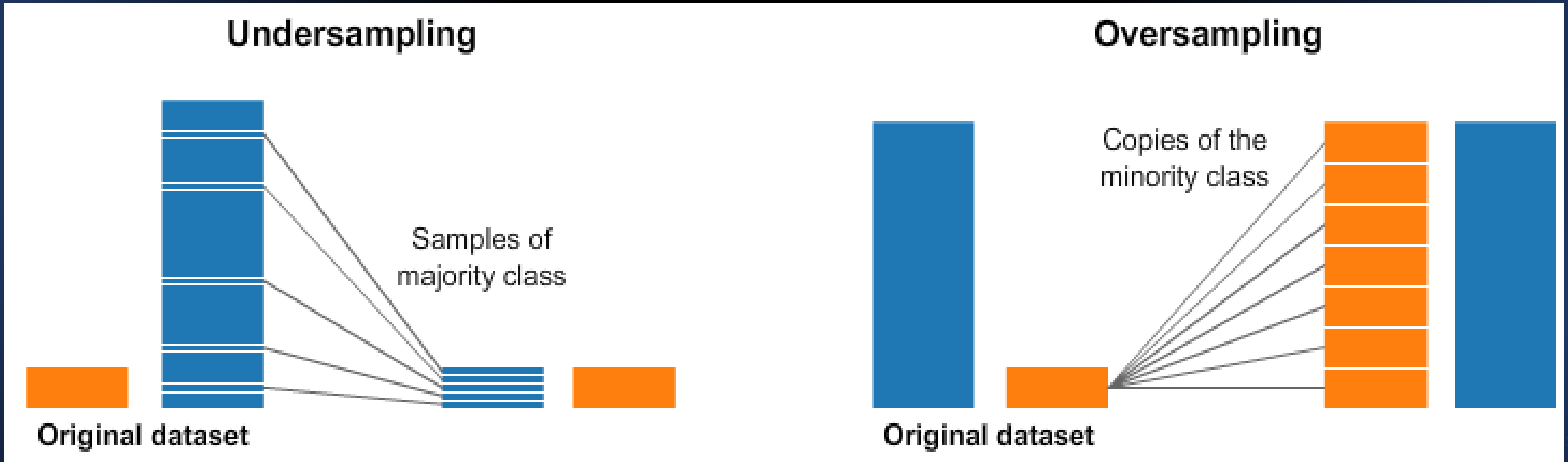


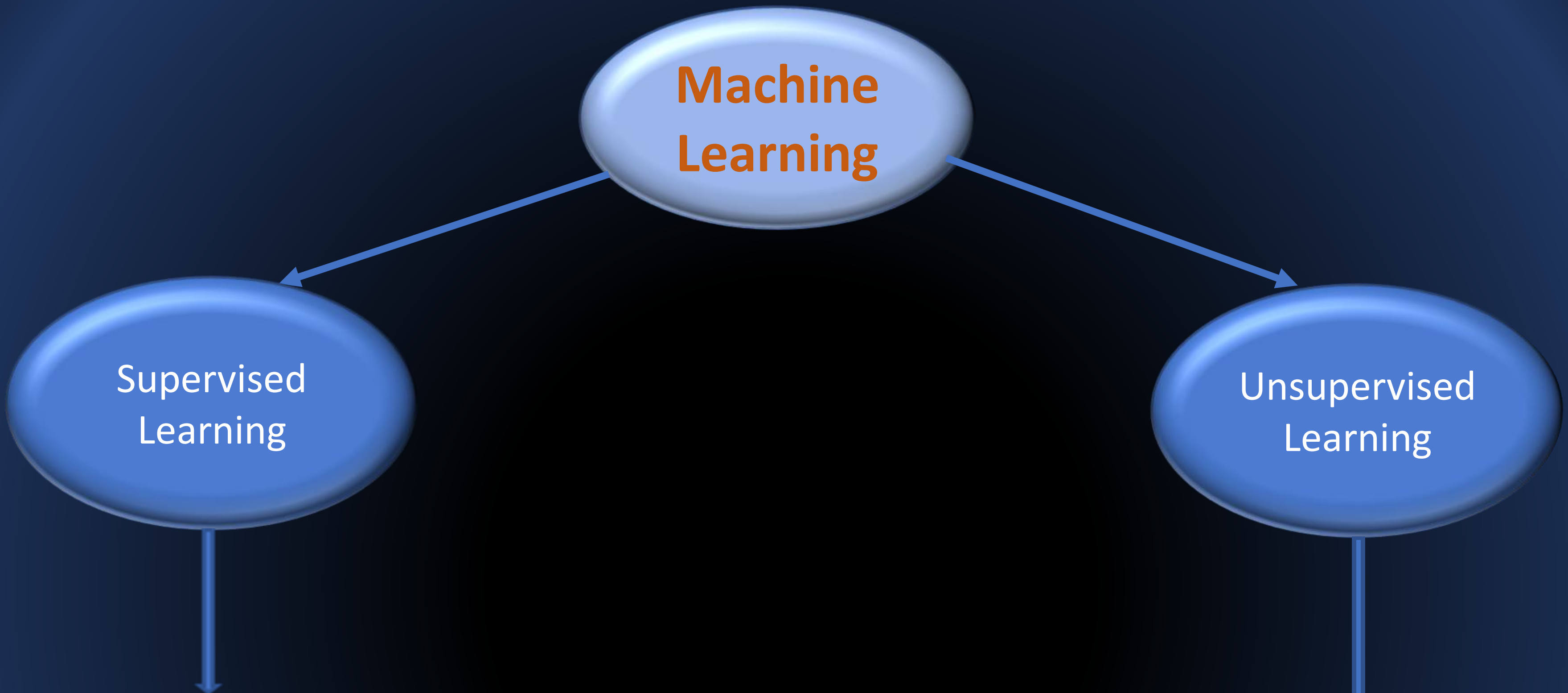
Data Pre processing

- Data we have used is highly imbalanced. Imbalanced data is a problem in model training which can result in high bias towards majority class.

- To handle the Class Imbalance we used following techniques

- Oversampling using SMOTE , ADASYN
- Undersampling using Random under sampler





Machine Learning

Supervised Learning

Unsupervised Learning

Models
Logistic Regression
Decision Tree
Random Forest
XG Boost
Stacking with KNN, Gaussian NB, Random Forest, Logistic Regression

Models
Neural Network Autoencoders
Clustering with t - SNE

Presented by

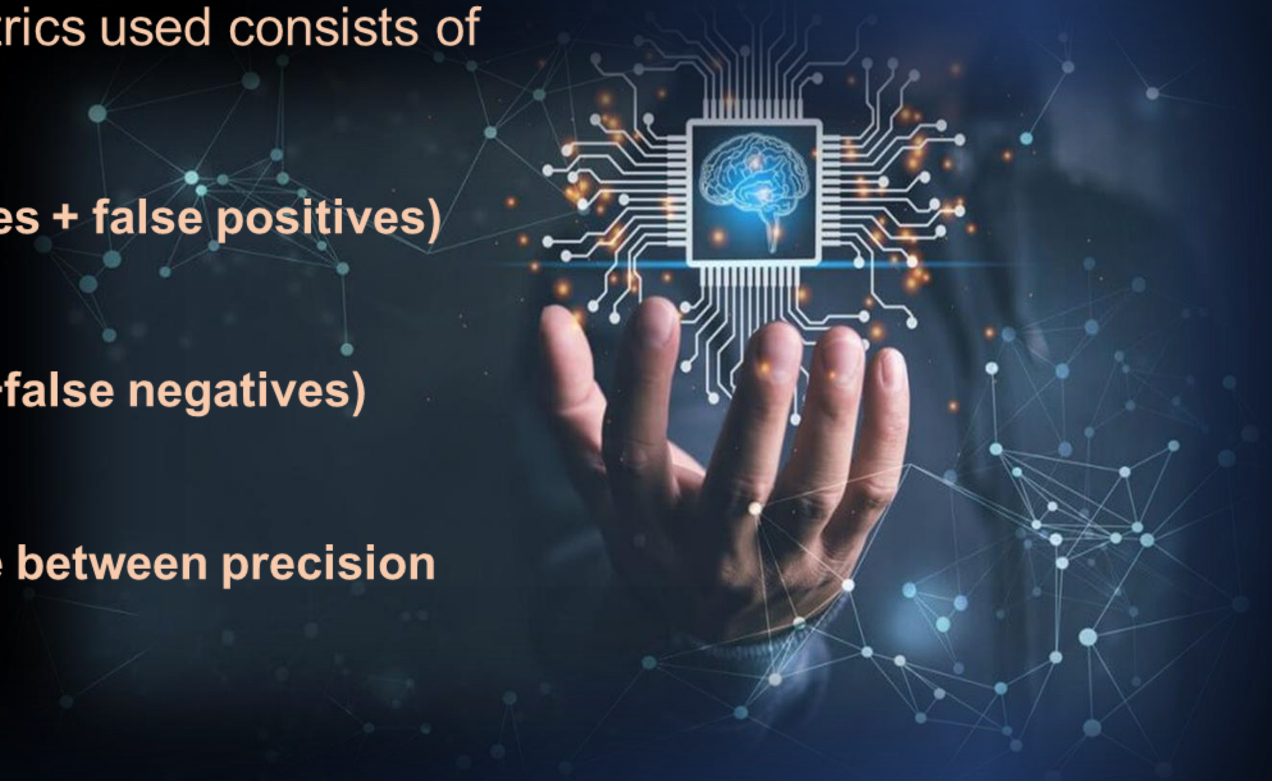
Sadvi Sandhya

Model Evaluation

Metrics

Classification Report: This is the key metrics used consists of

- **Precision : true positives / (true positives + false positives)**
- **Recall : true positives / (true positives +false negatives)**
- **F1 score : Takes into account a balance between precision and recall.**



Presented by

Sadvi Sandhya

Results

Methods		Models	Accuracy	F1 Score(Fraud)
Supervised Machine learning	Oversampling With SMOTE	Logistic Regression	97%	10%
		Decision Tree	98%	11%
		Random Forest	99.9%	87%
		Xgboost	99.9%	75%
	Oversampling With ADASYN	Random Forest	99.9%	86%
		Xgboost	99.9%	74%
	Under Sampling	Logistic Regression	95%	7%
		Random Forest	97%	10%
	Hyper tuning and SMOTE	Random Forest	99.9%	87%
		XGboost	99.9%	88%
Deep Learning	Neural Network Autoencoders	Autoencoders	99%	73%
Stacking	Base	KNN,GaussianNB, Randomforest,Logistic Regression	99%	84%
	Oversampling With SMOTE	KNN,GaussianNB, Randomforest,Logistic Regression	99%	84%

Presented by

Sadvi Sandhya

Conclusions

Annual Business Loss without ML models GBP 5.5M

With our Best performed model, we reduced the business cost to GBP 980/day

Annual Business loss minimised to GBP 356000

Annual cost saving with ML models GBP 5M

Presented by
Sadvi Sandhya

Future work

- To try Generative Adversarial Networks for Improving classification effectiveness
- To run the models in real time as additional Fintech data would provide more insights on the applicability of existing algorithms.

THANK YOU!

Presented by

Sadvi Sandhya