**Second Group:**

Penny Scams vs Non Penny Scams

Check out the false positive and false negative costs for the presentation Javelin Strategy Research

Choice of Performance Metric (F1 Score not suitable meh?)

Elaborate more on the SMOTE (SMOTE plus Easy Ensemble for cleaning process)

Model Pipeline (SMOTE, Easy Ensemble Training using ensemble of XGBoost)

Explain why XGBoost is used for the hackathon? (Parallelisation, Continued Training, Ability to identify feature importance) 🡪 continued training requires more data points to be introduced

Recall and Precision vs Ensemble Threshold ( Optimal Recall, Optimal Precision, Optimal Ensemble the ensemble threshold)

Peeny Frauds and Non-Penny Frauds (Three Reasons)

* Ecommerce False Positive
* Card holder Behaviour and Habit

Adaptibily as a Strength:

Business Implications and Recommendation (have four courses)

Have a decision process:

* Predict Positive + Predict Negative
* By Value (Low amount, med amount, high amount)
* By Confidence (p>0.95 etc)
* Course of Actions (Decline or Approve, follow up or SMS)

Limitations of Further Research:

* Anonymisation of Features
  + Using behavioural analytics
* Lack of Cost Visibility
  + Frustration and intervention costs (need more accurate cost details)
* Non-representative data
  + 40 hours is not sufficient

Current model is : recall and precision are furthe

Future model is : recall and precision are closer

Third Team:

EDA:

* Show normal distribution on the whole slide
* Consider removing the seconds after reference time
* Dealing with Missing Values (seems to be missing at random 🡪 use mice packages)
* Remove outliers 🡪 Performance decreases + info loss

Metrics:

* Which metrics used to compare model (Recall)
* Metrics used to tune the parameters (F1 Score)
* Choose baseline models (logistic regression)

Models:

* Explain why decision tree is chosen (base classifier) + show caveats
* Decision Tree:
  + Class Weights
  + Choose Max Depth (Tree Pruning + plot the graph)
* Ensemble Methods:
  + Random Forest (Bootstrap Aggregation Methods):
    - Explain the algo
    - Why decision tree chosen + caveats
    - No of estimators + max depth
  + Gradient Boosting :
    - Sequential ensembling method
    - XGboost chosen (boosting) (benefits + caveats) 🡪 long run time
    - Parameters: learning rate + No of estimators + max depth
    - Panalising to check the xgboost
* Validation set to tune parameters

Model Selection:

* XGboost is best

Business Solution:

* Do a graph (show inputs, xgboost model, outputs and iterative process)
* Customer experience (flat the fraud 🡪 sms text messages, check if it is authorizes transaction, customer report it as fraud)

**From Mentor**

Things to improve on:

* Models:
  + Parameter Tuning need to improve on
  + Models improvement
* Business Solution:
  + Need to propose the rules + integrate everything
  + Iterative Process

**Fourth Group:**

* Feature 23 is noisy (near uniform spread) 🡪 lead to overfitting
* Evaluation Metrics:
  + F-beta Score(able to assign scores) + AUC score
* Data Prep:
  + Explain over sampling and undersampling
* Model Training:
* - Check when will the model be diverged for training error and test error?
* Explain Cross Validation

Business Solution

* Attention Required
* Importance or Value of Customer
* Amount of Transactions
* Threshold set for probability of fraudulent activities

Explain why certain models are not used and what models are used instead:

* For neural network: good to use neural network due to scalability
* Neural network tend to overfit on small dataset (200 k is quite small) + how to prevent maybe using regularization

Fifth Group:

Seconds after Referencing Time-

* Can redo the visualization

Models Selection:

* SGDClassifier with loss function SVM
* SVC
* Can consider plotting confusion matrix
* Smoteenn (instead of SMOTE)
* WE NEED TO TAKE INTO account of scalability when presenting the solution!
* Check before and after Hyperparamter tuning

**Last Group:**

Parameter Tuning:

* Max Features

Data Resampling:

* Use k-fold cross validation for under sampling method
* Use SMOTE for oversampling methodology (shld do under-sampling first then over-sampling)
* Check out the cross validation score, false negative rate, false positive rate for both under sampling and over sampling

Others:

* Rule Based Fraud System
* Explain why speed of model is important

DO WE NEED TO USE LDA??

EVALUATION METRICS:

* Final score is a combination of score on everything
* NEED to be succint and clear on data processing process (Need to be CLEAR)
* Have screenshot of the slides on the codes (highlight the codes, can part of the feature)
* Need to say what evaluation metric is used (need to explain why and be CONSISTENT)
* Business Solutions (need to learn how to integrate everything tgt)
* Model (explain 1 or two slides that we do modelling and parameter optimization)
* FINAL (no feedback + more time, 25 minutes presentation + any amount of slides is sufficient)