NAÏVE BAES

1. Why lower amount?
   1. Answered well
2. Why feature 24?
   1. Answered well

TO DO

1. **Geospatial must be part of the limitations - J**
2. **Give light on how you prepped the data? - J**
3. **How you confirmed that it’s clean - J**
4. **Add more parameter tuning**
5. **Need to improve on business solution proposal**
6. **Submit multiple models, based on Company input (prioritise f1, auc, recall, or precision), select the best model**
7. **Why Random Forest – Can say it gives rules and we can apply that to business. IMPORTANT**
8. **2 types of answers: Prelim and Detailed**
9. **Client might have different number of people to investigate: Choose if they want top 5% of frauds or top 50%**
10. **Ask judges what they want then and there. REcmmend model based on htat**

# Jingle Jangle

Penny Fraud vs Non-penny Fraud

* Penny fraud is <$5

Misclassification cost breakdown

* Less revenue, cost of intervention etc

There’s a lot of focus on cost breakdown.

* Judges really nod when he spoke about cost of false positives!

Data resampling

* EasyEnsemble + SMOTE
  + Every subset contains all fraud cases with a sample of non-fraud cases

XGBoost

* Ability to identify feature importance
* Parallelisation
* Continued Training

Evaluation Criteria

* Focused on recall

ADAPTABILITY AS A STRENGTH

* First spoke about recall – their model is tuned to improve it
* Suddenly decided precision is better – model doesn’t need to be retrained, just change parameter. BEAUTIFUL

Business Implications AND Recommendations

* Added a course of action based on probability and fraud value

MIC DROP

* Linked last slide to first and said they saved a lot of cost

FEEDBACK THEY GOT

Very informative. Lots of research done. Very close to how the biz works.

Question on XGBoost – what algos are we using? XGBoost is apparently good for a particular dataset.

His answer: XGBoost is able to use parallel processing such that runtime is shorter cz the data keeps coming in.

Q: XGBoost also requires continues training. Did you manage to do that?

A: No. No additional data coming in that’s why

Q: Did you use any algo with that?

A: NO

# Net Surfers

Metrics

* One metric to compare between models
* One metric to tune the same model

Also used XGBoost

**Any point where parameters are inputted by the company? – As a bank, they may want to tune how to catch parameters**

Their thought process in choosing which model was nice. We need to justify this more.

**Didn’t get logic as why they chose Random Forest and XGBoost??**

3 Heavenly Kings

Used keras and Neural Net

Dropped null rows of features 5,6

Dropped feature 23

Used F-beta score and AUC

Biz soln

* Process flow and all
* Profile users

Feedback

* How much time on data processing?
* Why NN?
  + NN can be scaled down or scaled up
* There was no mention of limitations of dataset
* How would you prevent overfitting?
  + One possible way is regularisation

# Team Whack and Hack

Graphs are very colourful

* Probably cz both women

Models

* SGDClassifier with loss function SVM
  + Not as accurate as SVC
* SVC
  + Runtime is very high
* Random Forest Classifier
  + Better

Metrics used

* Recall, F1, Precision

Data Balancing

* SMOTEEEN
  + SMOTE + Edited Nearest Neighbour to clear it out

Model

* RandomizedSearchCV
* Random Forest

Feedback

* Good explanation of why this model
* Good biz description
* How can you solve less data issue?
  + Cross validation

# Does this even matter

Dropped the rows that contain NA values

Models tested

* Logistic regression
* SVM
* Naïve Bayes
* Random Forest
* LDA

Cross validation

Used GridSearchCV

Conduct SMOTE on cross-validated training set

Fit model with cross-validated training set

Test model accuracy on cross-validated testing set

Really had a lot of cross validation results man

**Feedback:**

RF uses a lot of computational power and is quite complex. Did you face those issues?

* No, blab la

Why tf did you use RF man. Just answer meeeeeeeeee