SO MANY MODELS SO LITTLE TIME

THINKFUL SUPERVISED LEARNING CAPSTONE

KEVIN LAM

BACKGROUND

- Found on Kaggle: https://www.kaggle.com/ntnu-testimon/paysim1
- Synthetic financial dataset created for fraud detection generated by Paysim.
- Fraud- fraudulent mobile money transactions where an agent attempts to gain access to a customer's account and empty the funds by transferring to another account and cashing it out of the system.

GOAL

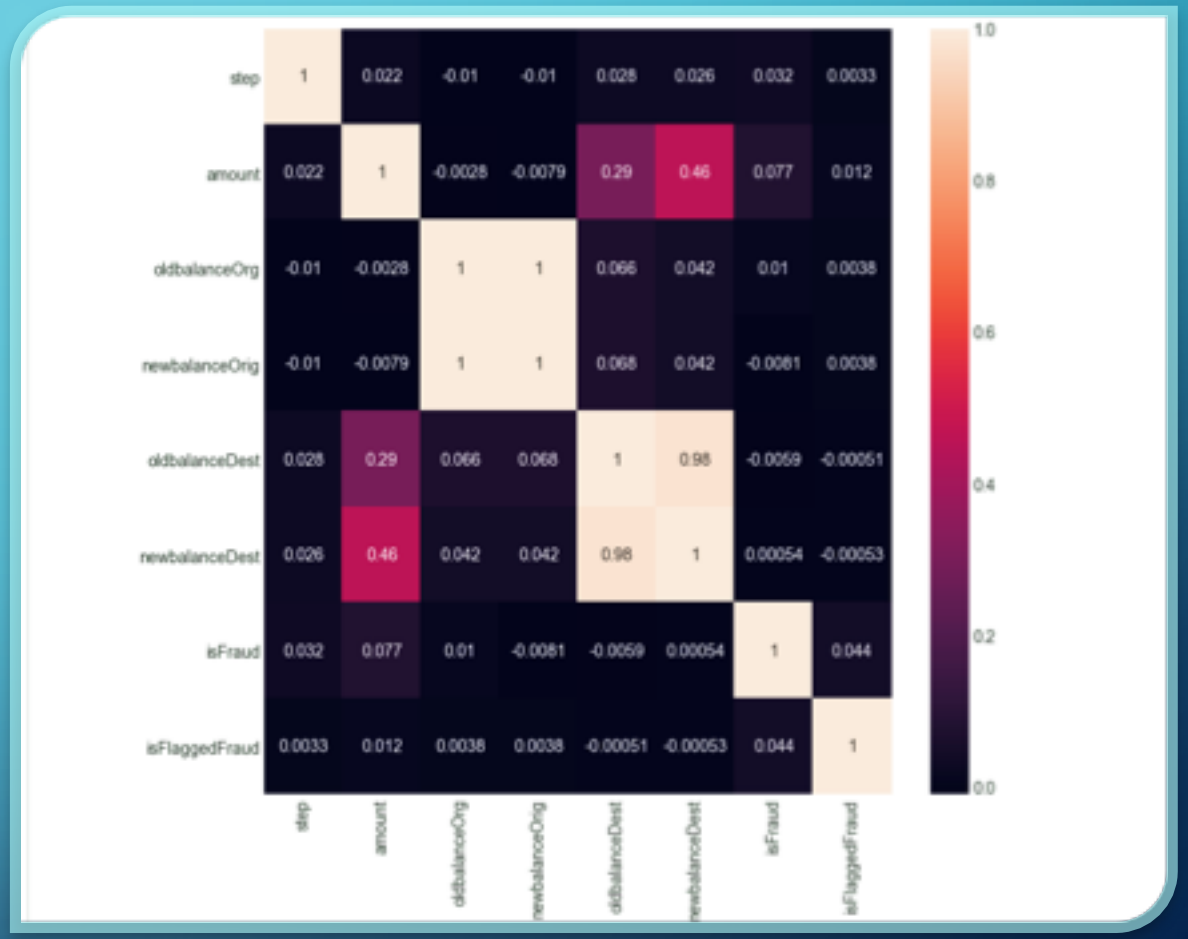
- Two type of models:
  - I want to testify, I need to CLASSIFY
- Find the best model to accurately label data as fraud.
  - BEST – Not only fast, but accurate too.

DATA EXPLORATION

- 11 features in the dataset:
  - Categorical: Type, nameOrig (customer who started the transaction), nameDest (recipient of the transaction), Fraud, Flagged as Fraud (isFlaggedFraud)
  - Step, Old/ New Balance Owner, amount, Old/New Balance Destination
- What to do with the categorical data?
  - Get dummies for the type column.
  - Drop the names.

DATA EXPLORATION CONT'D

- Created a heat map to see if any of the variables had any sort of unexpected multicollinearity.
- Dataset had 6,362,620 rows originally.
- There weren't any missing values.
- Suffering from class imbalance:
  - Only 8213 out of the 6,362,620 were fraud.



CLASS IMBALANCE

- Class Imbalance, so what?
  - Most of the data is non-fraudulent, not balancing the data will skew any model and make it 99% accurate even though it isn't.
  - Resample: 100,000 cases of true and false.

BUT WAIT… THERE'S MORE

- Feature Selection Method
  - SelectKBest vs Principal Component Analysis
    - SKB removes all but the highest scoring features.
    - Scoring in this case is determined by the f_classifier aka using the f-test to determine if there is any statistical significance in the sample variance of the dataset.
    - Principal component analysis transforms the all the features in the dataset and gives entirely new features that are independent from each other.
    - I chose to use 4 Principal Components.
  - Grading Rubric:
    - Accuracy, Runtime
    - AUC

MODEL 1: GAUSSIAN NAÏVE BAYES

- Why Gaussian?
  - Because there were initially more continuous variables before I got dummies for the categorical type data and that's the only way I can incorporate them unlike Bernoulli and Multinomial NB.
- SKB Results:
  - Accuracy: 77.90%|| AUC: 15.44% ||Runtime: Fast
- PCA Results:
  - Accuracy: 83.43%|| AUC: 37.62%||Runtime: Fast

MODEL 2: KNN

- Had to prep this when using SKB by normalizing distance in features.
- SKB Results:
  - Accuracy: 99.37%|| AUC: 0.51% ||Runtime: Fast <3 mins
- PCA Results:
  - Accuracy: 98.84%|| AUC: 37.67%||Runtime: Faster <30 seconds

- Overall: Both models overfitted.

MODEL 3: DECISION TREE

- SKB Results:
  - Accuracy: 98.76%|| AUC: 48.95% ||Runtime: Fast
- PCA Results:
  - Accuracy: 93.13%|| AUC: 41.22%||Runtime: Faster
- Overall: Although SKB looks superior, the possibility of the model overfitting is quite high so I would rather go with the PCA model here.

MODEL 4: RANDOM FOREST

- SKB Results:
  - Accuracy: 97.46%|| AUC: 48.25% ||Runtime: Slow, ~2.5 hrs
- PCA Results:
  - Accuracy: 93.27%|| AUC: 42.30%||Runtime: Moderate, 18 minutes
- Overall: Both used maximum depth allotted in GridSearchCV. PCA model needed 25 more trees. PCA model less likely to have overfitted.

MODEL 5: LOGISTIC REGRESSION

- Used the Ridge Regularization due to the amount of rows. Had to prevent the overfitting of the coefficients.
- SKB Results:
  - Accuracy: 90.73%|| AUC: 43.22% ||Runtime: Fast, 2 mins
- PCA Results:
  - Accuracy: 83%|| AUC: 37.85%||Runtime: Faster, 9.2s
- Overall: SKB definitely is the superior model this time. The time difference is negligible.

MODEL 6: GRADIENT BOOSTED MODEL

- SKB Results:
  - Accuracy: 99.85%|| AUC: 0.12% ||Runtime: LONG, 8 hours
- PCA Results:
  - Accuracy: 99.33%|| AUC: 22.84%||Runtime: Faster, 5 hours
- Overall: Both are greedy models. Used the maximum amount of parameters to achieve the best score. Tried using a learning rate to inhibit overfitting but it wasn't effective.

WHERE IS SVM?

- It took too long to run.
- Even tried to use a linear kernel but to no avail.

CONCLUSION

- In production, if only one model is allowed to be used to classify whether an activity is fraudulent or not:
  - PCA Random Forest is the way to go. Superior accuracy with a moderate calculation time.
  - Runner up: SKB Logistic Regression given its superior run time and moderate accuracy.

QUESTIONS?

Find the notebook at: https://github.com/lamka/Thinkful-Unit-3-Capstone/blob/master/Unit%203%20Capstone%20Project.ipynb