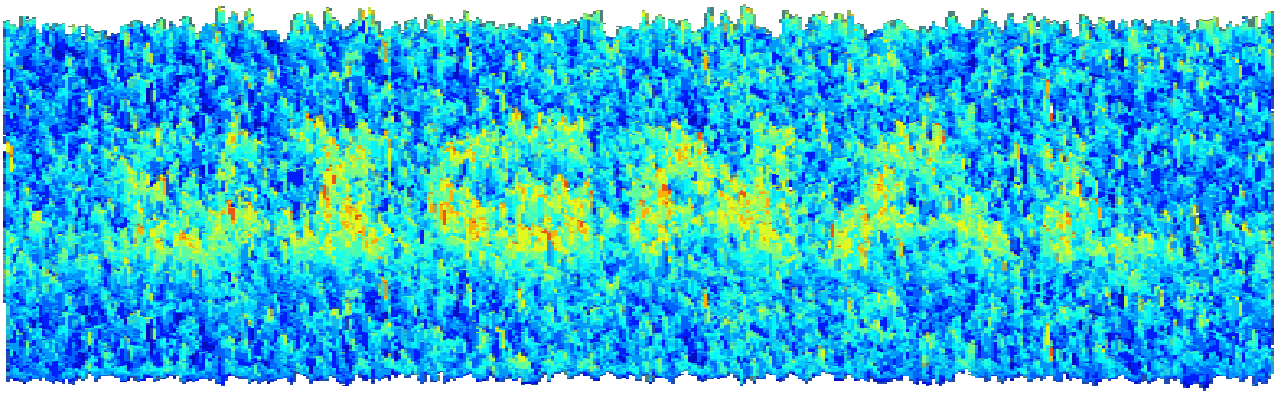

Financial Fraud Dataset - Basic Pipeline

NOTE: READ THIS DOCUMENT THOROUGHLY FOR SUCCESS

Background

Throughout our exploration of various supervised learning algorithms, a common theme that you all might have detected (no pun intended) is the concept of extracting **some signal from a noisy dataset.**

This often applies to datasets where our class of interest (***fraud***, ***basketball upset, etc)*** is in the “minority” class of an imbalanced dataset.

This problem often occurs in the detection of **financial fraud**. While we can assume that *most* transactions are credible, classifying every new sample as **non-fraudulent (0)** will **miss every single fraudulent case**.

Within this project, we will take a look at a synthetic dataset of bank transactions to see which strategies we can take in order to successfully **capture as many fraudulent transactions** as possible, while also **minimizing false positives**.

This dataset contains a mix of discrete and continuous variables. As you review this list of columns, **consider which columns might not be relevant to your analysis**:

- **Step**: *A unit of time that represents hours in the dataset. Think of this as the timestamp of the transaction (e.g. hour 1, hour 2, … hour 534, …)*

- **Type**: *The type of transaction*

- **Amount**: *The amount of money transferred*

- **NameOrig**: *The origin account name*

- **OldBalanceOrg**: *The origin accounts balance* ***before*** *the transaction*

- **NewBalanceOrg**: *The origin accounts balance* ***after*** *the transaction*

- **NameDest**: *The destination account name*

- **OldbalanceDest**: *The destination accounts balance* ***before*** *the transaction*

- **NewbalanceDest**: *The destination accounts balance* ***after*** *the transaction*

- **IsFlaggedFraud**: *A "naive" model that simply flags a transaction as fraudulent if it is greater than 200,000 (note that this currency is* ***not*** *USD)*

- **IsFraud**: *Was this simulated transaction actually fraudulent? In this case, we consider "fraud" to be a malicious transaction that aimed to transfer funds out of a victim's bank account before the account owner could secure their information.*

Within this project, you will be creating a comprehensive machine learning pipeline that satisfies the **patterned steps** of a classic machine learning project. You will:

- *begin with* ***hypothesis formulation through EDA****,*
- *complete* ***data cleaning & pre-processing****,*
- *and conclude* ***with model generation and a report.***

<u>Instructions</u>

The following is a list of **expected notebooks** that should be included in your project:

1. **Initial EDA**
    - Your project should begin with a notebook where you perform **univariate**, **bivariate**, and **multivariate exploratory analysis**.
    - Be sure to create **relevant graphs** that will help you **formulate a hypothesis**.
    - Keep in mind that we are not only interested in the relationships between our **predictors and our target variable**, but we are also interested in the **relationship amongst predictors**.

2. **Data cleaning, wrangling & pre-processing**
    - After completing your EDA, you should move forward with creating a notebook where you will **clean and wrangle your dataset.** This might include dropping **null values**, removing **unnecessary columns**, removing **outliers**, and **potentially fixing incorrectly formatted data**.
    - Save this dataframe as a new csv file to be used in the next step.

3. **Model creation, hyperparameter search, and model evaluation**
   - Once we've created this pre-processed dataset, we will then create a notebook where we will create **train test splits** and implement a RandomForestClassifier or a GradientBoostingClassifier *(the choice is yours!)*
   - After training this initial classifier and viewing its accuracy measures, we will then move forward with **hyperparameter tuning** via GridSearchCV or RandomizedSearchCV *(again, the choice is yours!)*
   - Upon finding **optimal hyperparameters**, we should re-train our model using these hyperparameters, generate predictions for this new model, and output the subsequent F1 score for this classifier.

4. **Report**
   - To conclude this project, we should answer the following 5 questions in a **separate document attached to your project:**
     i. **Which insights did you gain from your EDA?**

     ii. **How did you determine which columns to drop or keep? If your EDA informed this process, explain which insights you used to determine which columns were not needed.**

     iii. **Which hyperparameter tuning strategy did you use? Grid-search or random-search? Why?**

     iv. **How did your model's performance change after discovering optimal hyperparameters?**

     v. **What was your final F1 Score?**

FAQ

1. *This project seems really ambiguous, how should I start?*

Part of being a successful technologist is figuring out how to start a project yourself. Oftentimes, we do not have frameworks to base our projects on, except **for our own work**.

Therefore, we recommend that you **look back to your previous TLABS and in-class labs to see which code could potentially apply** to each part of this project.

Looking back to our first TLAB (Rental Price Prediction), it looks like we have some template code for at least our EDA.

2. *How am I supposed to run my analysis if I don't include my dataset in GitHub?*

   Keep in mind that just because something is not pushed to GItHub, does not mean that you will **not be able to use it on your local machine** (even if it is in your cloned git repository!)

3. *I don't know how to make a RandomForest/GradientBoosting model. Where do I look to find this out?*

   We provide links to both models. Check out the example code to get started. Also we recommend that you use Google to see how other data scientists implement these learning algorithms.

4. *How do I implement hyperparameter tuning?*

   We provide links to both Grid & RandomizedSearch objects. Check out the example code provided, as well as your code from the previous weeks labs for ideas.

5. *Which model/hyperparameter tuning strategy do I choose?*

   This is **your** executive decision to make. Better yet, why don't you **try out both** and see which one performs better?

6. *Which detail should my report include?*

   Check out our slides from 11/29 on data reporting for some insight on how we expect you to report your findings.

Submission

Submit a link to your repository with all required **notebooks and write ups by 2/22**

For an example of how this machine learning project could look like, check out the following repositories:

- Fish Toxicity
- Laptop Price Modeling
- Student Performance Prediction

We are primarily interested in seeing which ideas you took away from the previous weeks of supervised learning. **Hesitate to copy ChatGPT code, you will not receive accurate feedback by providing us with LLM generated code. We want to see your work.**