Capstone Project
Real / Fake Job Posting Prediction

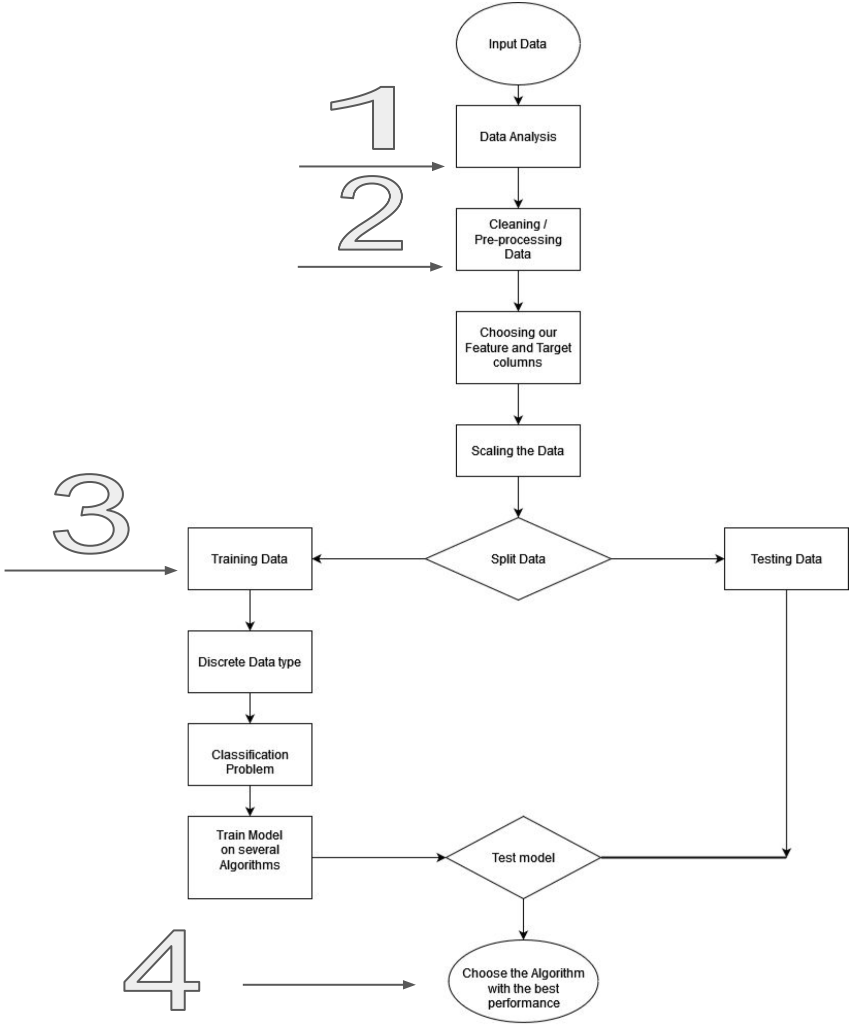
Purpose

Job search resources are being used by hackers as a method for phishing attacks in order to get personal information resulting in identity theft and money scams.

This Project aims to create a machine learning model that can accurately identify fraudulent jobs postings before they are advertised to the general public

Processes

1. Data Analysis
2. Cleaning Data
3. Training Data
4. Validate data with different algorithm
5. Check the Result


Input Data
1
Data Analysis
2
Cleaning / Pre-processing Data
Choosing our Feature and Target columns
Scaling the Data
3
Training Data
Split Data
Testing Data
Discrete Data type
Classification Problem
Train Model on several Algorithms
Test model
4
Choose the Algorithm with the best performance

Data Analysis - Source

- Dataset obtained from Kaggle
- Database Integration with Postgres & AWS RDS
- 17838 Rows by 18 Columns
- Target Feature in dataset it “Fraudulent”

Data Analysis/Cleaning - Dropped Features

The following explains analysis of features that were dropped from the dataset before processing

- Job ID is an identification column and has no relation to the target Feature
- Salary_range was majority null values
    - <graph showing number of null values>
- title contained majority unique values
    - <graph showing number of unique values>

Data Analysis/Cleaning - Features and Target

- Dataset has 17 features
    - 4 features have large amounts of text and provide description of the job posting and the company posting the job
    - NLP using the NLTK library will be used to remove punctuation, stopwords then lemmetize the text before tockenizing and then running Term Frequency (TF-IDF) to determine the relevancy of keywords
    - "company_profile", "description", "requirements", "benefits"

<Graphs showing the 4 features above with number of null values for each>
<graph showing the 4 features above with the how many were fraudulent vs how many were not>

Data Analysis/Cleaning - Features and Target

- Target Feature in dataset it “Fraudulent”
- Dataset has 17 features
    - 3 features identified as text based ordinal data types with specific ordered groupings
    - LabelEncoder used to convert these features to numeric form for the machine learning model
    - 'employment_type','required_experience','required_education'

<Graphs showing the 3 features above with number of null values for each>
<graph showing number if unique values for each feature>
<graph showing the 3 features above with the how many were fraudulent vs how many were not>

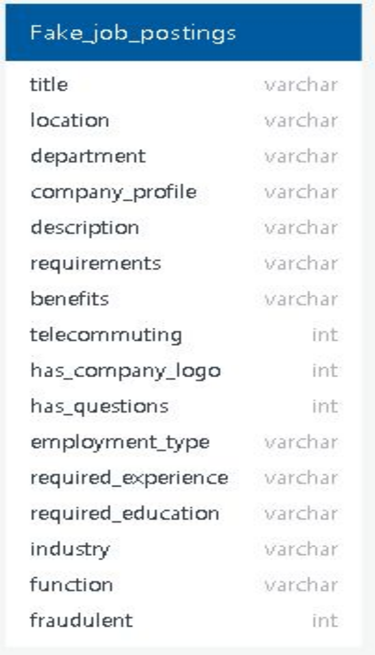
Data Analysis/Cleaning - Features and Target

- Target Feature in dataset it “Fraudulent”
- Dataset has 17 features
    - 4 features identified as text based nominal data types with unordered groupings
    - TargetEncoder used to encode in order to convert these features to numerical form based on the mean of the target to the count of each category
    - 'department', 'industry', 'function', 'Country'

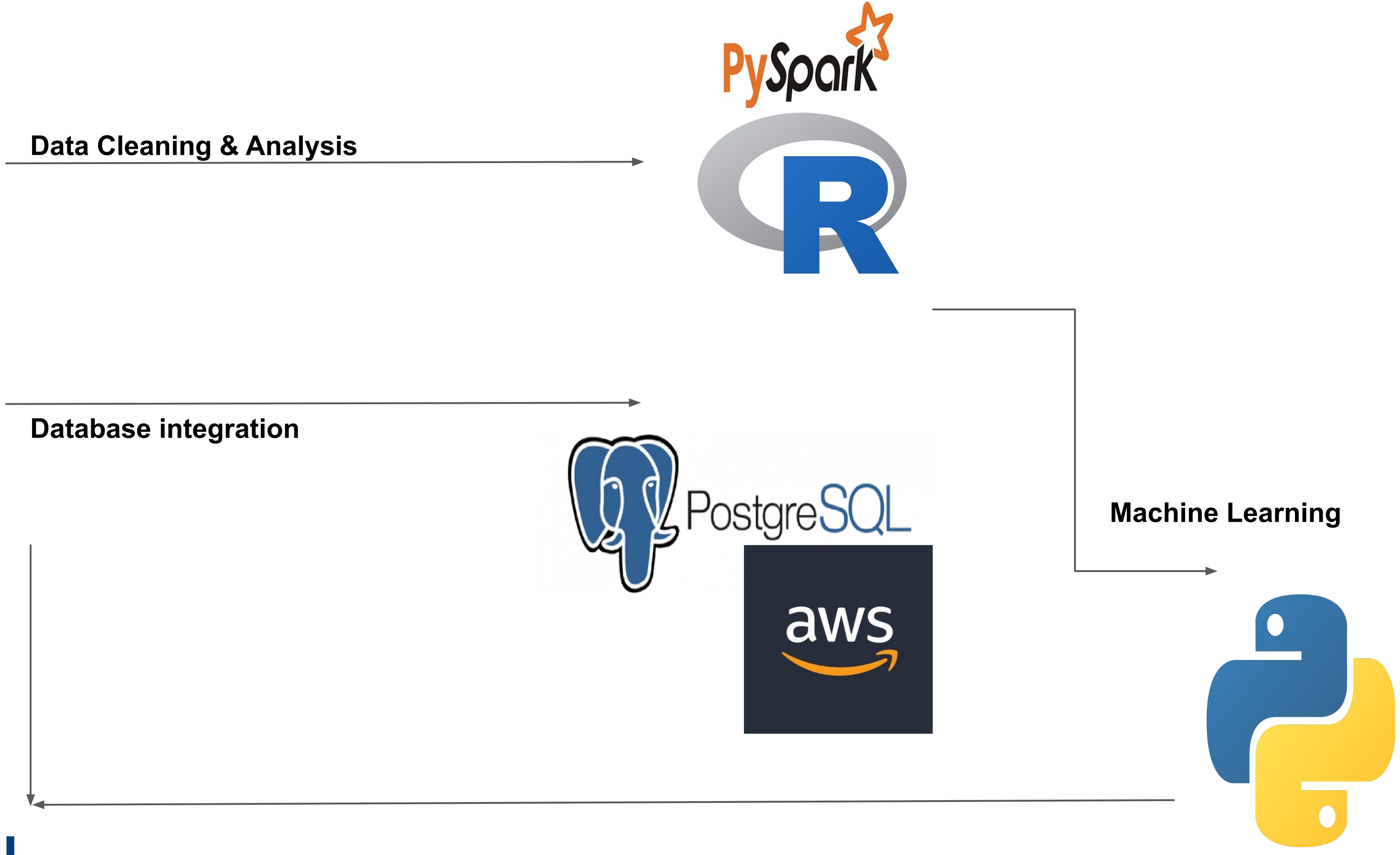
<Graphs showing the 4 features above with number of null values for each>
<graph showing number if unique values for each feature>
<graph showing the 4 features above with the how many were fraudulent vs how many were not>

Machine Learning

- Main analysis through machine-learning:
  - To predict fraudulent from job-posting data
  - Goal Accuracy: ()
- Supervised Learning (Target Column = Does Exist)
- Classification (Discrete Variable = Fraudulent vs Not Fraudulent)



Fake_job_postings	
title	varchar
location	varchar
department	varchar
company_profile	varchar
description	varchar
requirements	varchar
benefits	varchar
telecommuting	int
has_company_logo	int
has_questions	int
employment_type	varchar
required_experience	varchar
required_education	varchar
industry	varchar
function	varchar
fraudulent	int



Data Cleaning & Analysis

Database integration

Machine Learning

Visualization & Presentation

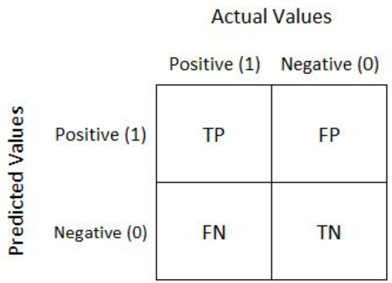
tableau

Tools that will be used to create the final dashboard

- Tableau will be used to create the final dashboard

  Once ML is set:

- Classification Model, Confusion Matrix will be displayed in the dashboard



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Final Dash Board Outline

- **Exploratory Data Analysis**
    - Basic Data analysis:
        - Proportion of Fraudulent vs Non-Fraudulent
        - Count analysis of columns: (Required Education, Industry, Fraud Rate)
- **ML Part:**
    - lightGBM + important features
    - Classification report and confusion matrix to assess the accuracy