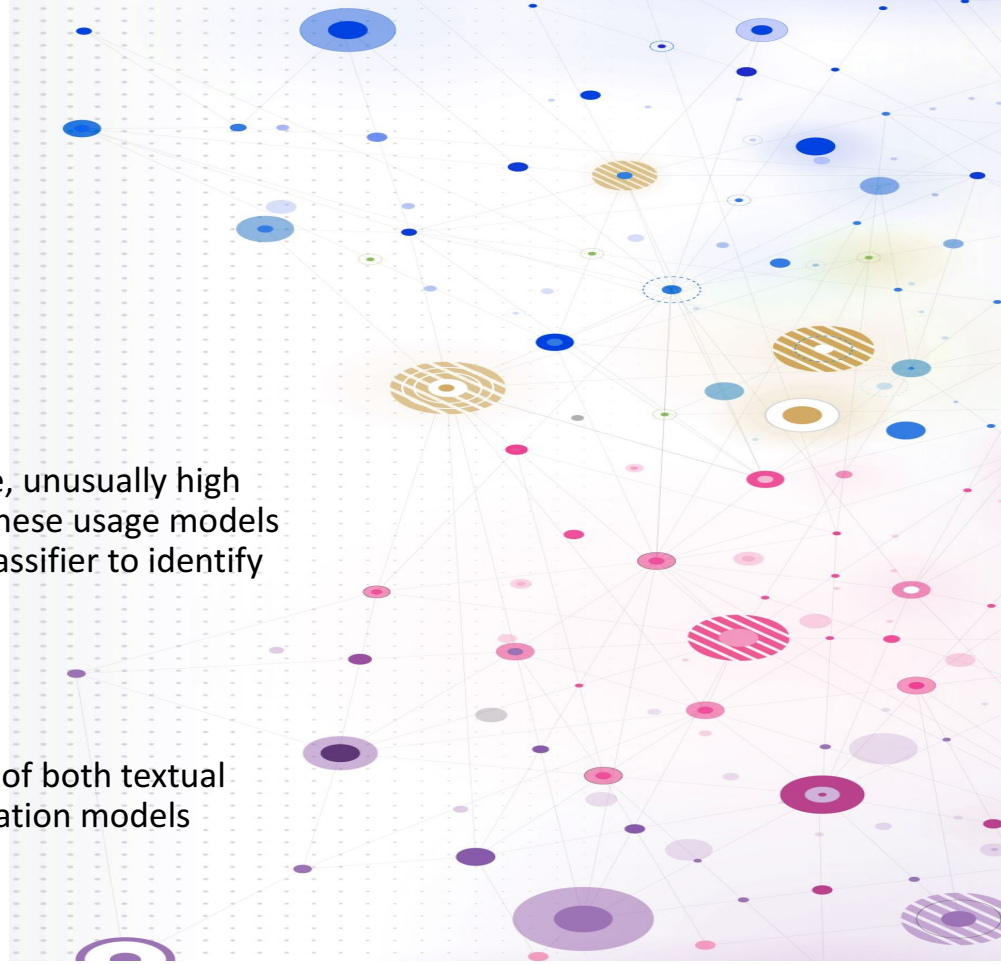
Aleksander Wojnarowicz(AW77438)

Yuqing Liu(YL110057)

Data Mining Final Project 2021



1.1 Motivation

In the past few years, people have used intuition to distinguish fake job postings. For example, unusually high salaries may suggest false job postings. Nowadays, big data technology allows us to process these usage models to release job data more reliably and identify fake data. The goal of our project is to train a classifier to identify fake or real job postings using functions such as salary range, benefits, Required_experience, Required_education, etc.

1.2 Dataset Description

This dataset contains 18K job descriptions out of which about 800 are fake. The data consists of both textual information and meta-information about the jobs. The dataset can be used to create classification models which can learn the job descriptions which are fraudulent.

Link: https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction

1.3 Description of Task

Our task is to bulid some models to predict whether the job is fake or not.

1.4 Descripe Target Variables

• fraudulent - the job is fake or not? (binary: 'yes', 'no’)

Features Explanation

Variables	Explain
Job_id	Unique Job ID
title	The title of the job ad entry.Most likely unique for each entry.
location	Geographical location of the job ad : Country, State, City
Department	Corporate department (e.g. sales).Most likely unique for each posting.
Salary_range	Indicative salary range
Company_profile	A brief company description
description	A brief company description
requirements	Enlisted requirements for the job opening.
Benefits	Enlisted requirements for the job opening.
Telecommuting	Enlisted requirements for the job opening.
Has_company_logo	True if company logo is present.
Has_quesitons	True if company logo is present.
Employment_type	True if company logo is present.
Required_experience	True if company logo is present.
Required_education	True if company logo is present.
Industry	True if company logo is present.

>>STEPS

1. Cleaning Dataset
2. EDA
3. Data Processing
4. Build the Model
5. Summary

Cleaning Dataset

Dataset Report: https://github.com/AleksanderWWW/Data_mining-Fake_job/blob/master/DatasetReport.7z

Cleaning Dataset

1.1 Removing Missing

Delete unnamed col

Delete *'telecommuting*、*has_company_logo*、*has_questions*、*fraudulent*' not 1 or 0

Unbalanced Dataset

1.2 Imputing Missing

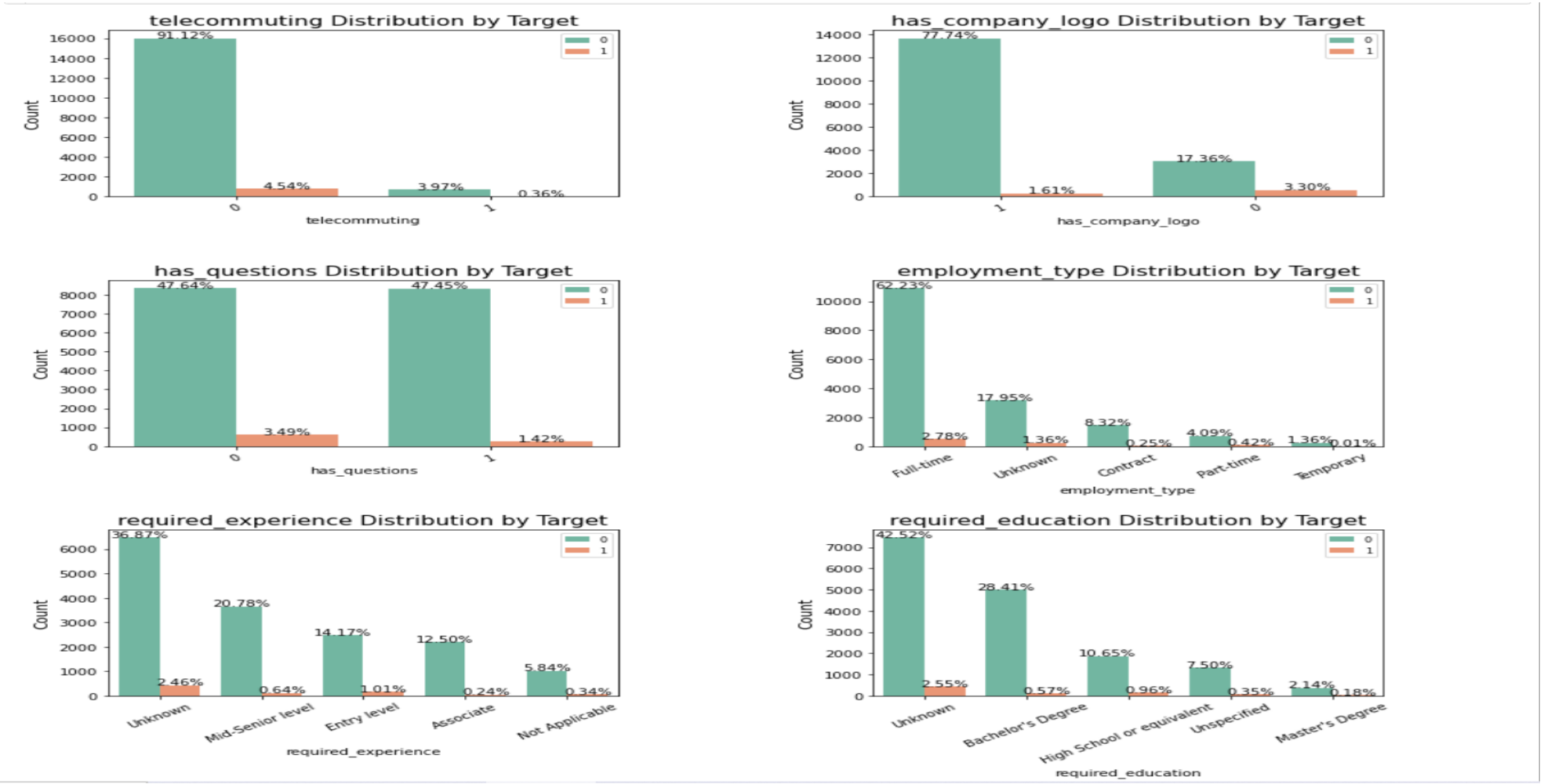
1.3 Handling Special Variable

Salary>> Max & Min

Convert numeric variable format from str to numeric

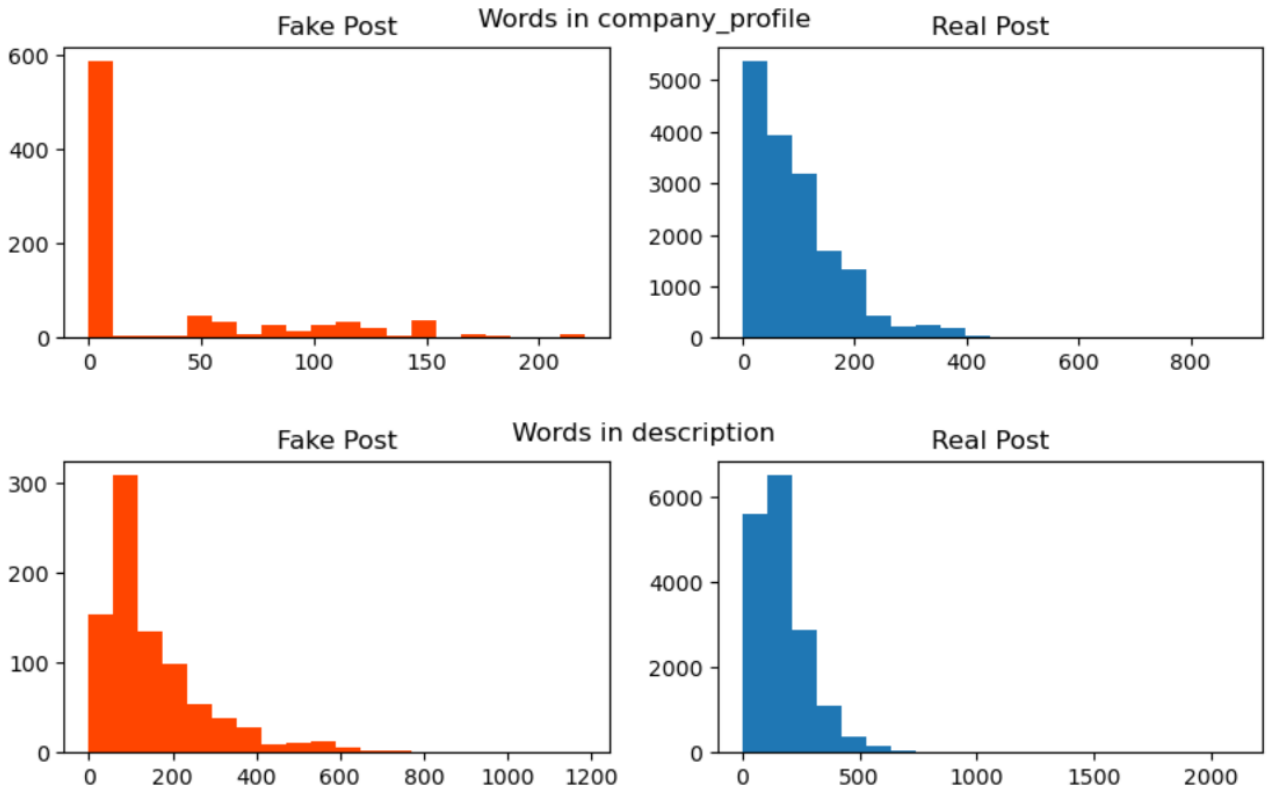
EDA

Education Related Variables



- **Real Posts**
  - ✅ Telecommuting
  - ✅ Company Logo
  - ✅ Employment Type
  - ✅ Required Experience and Education
- **Fradulent Posts**
  - ❌ Telecommuting
  - ❌ Company Logo
  - ✅ Employment Type
  - ✅ Required Experience and Education

Text information variables




Fake Post
Words in requirements
Real Post
0
100
200
300
400
0
2000
4000
6000
8000
10000
0
250
500
750
1000
1250
1500
Fake Post
Words in benefits
Real Post
0
100
200
300
400
500
0
2500
5000
7500
10000
12500
0
100
200
300
400
500
600
700

3. Data Processing

3.1 Key-Value

Descriptive related variables converted to length

3.2 One-hot encode

Ordinal variables

Discrete variable

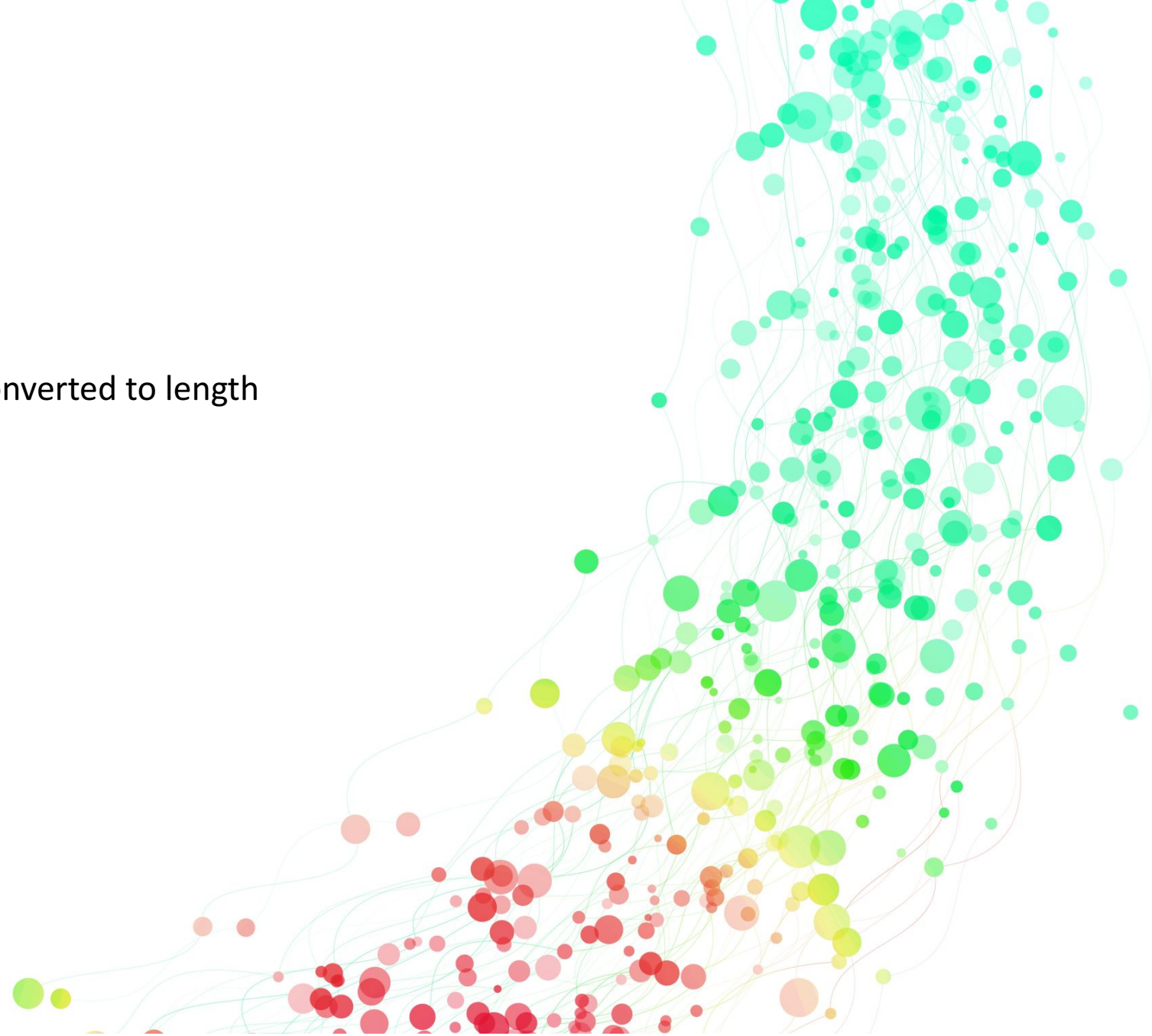
3.3 PCA

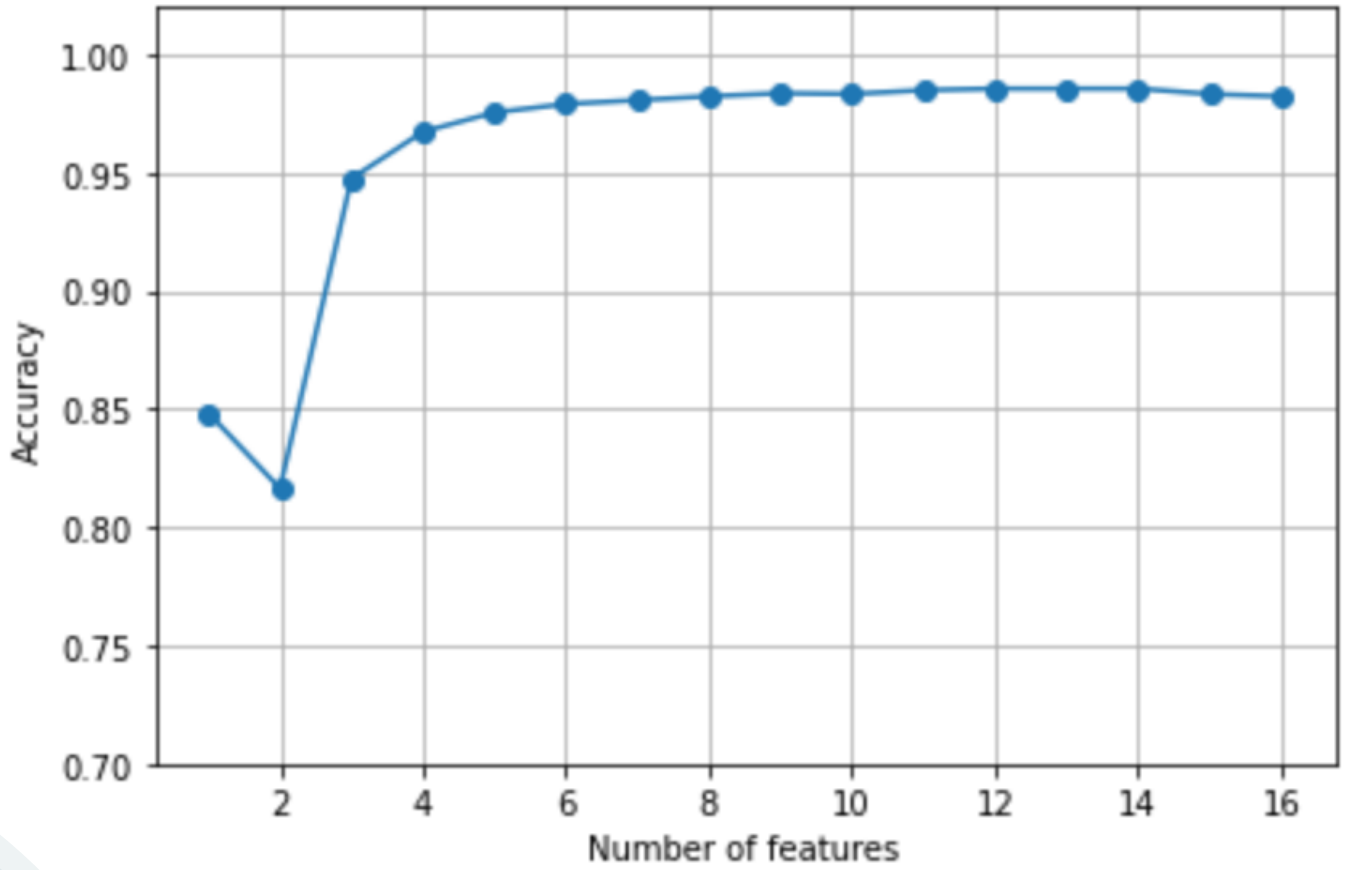
3.4 Standard Scale

3.5 Upsampling

3.6 Clone

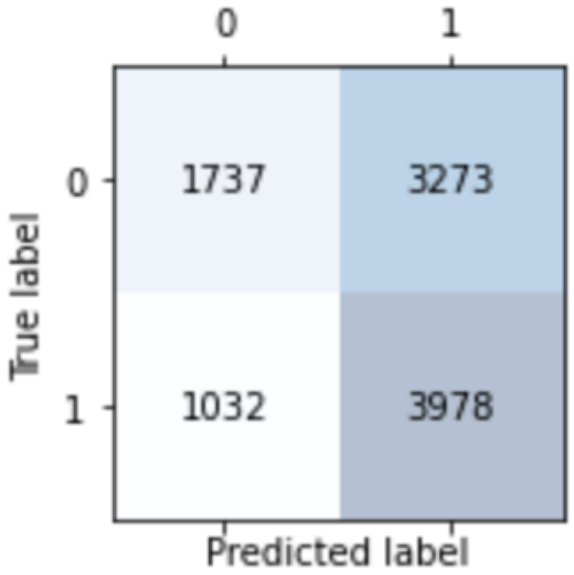
3.7 Feature Selection




1.00
0.95
0.90
0.85
0.80
0.75
0.70
Accuracy
2
4
6
8
10
12
14
16
Number of features

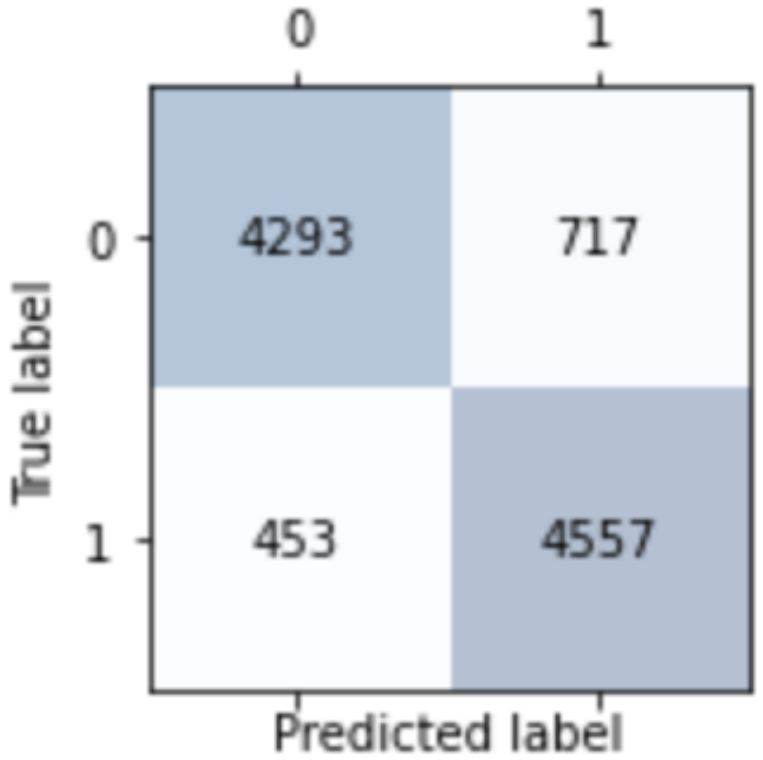
4. Build the Model

PCA_LR



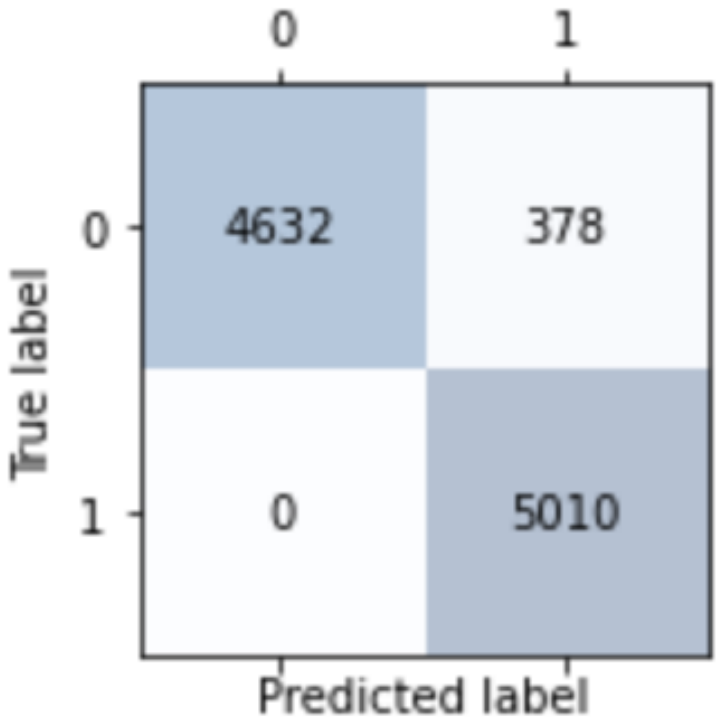
Precision: 0.627
Recall: 0.347
F1: 0.447

PCA_ Decision Tree



Precision: 0.905
Recall: 0.857
F1: 0.880

Bagging



Precision: 1.000
Recall: 0.925
F1: 0.961

Summary

- LR | 0.447
- DT | 0.879
- BA | 0.961

Thank you for your attention!