INTRODUCTION

NAME: VIVIAN KERUBO MOSOMI

Email: kerubomosmi7@gmail.com

Country: Kenya

College: Jomo Kenyatta University of Agriculture and Technology (JKUAT)

Specialization: Data Science

Problem Description

ABC Bank wants to predict whether a particular customer will buy their term deposit product based on their past interactions with the bank or other financial institutions. This will allow the bank to focus its marketing efforts on customers who are more likely to purchase the product.

GitHub Repository Link

https://github.com/Vee2002/DataGlacier_Internship/tree/vc/Data%20Glacier

Dataset Overview

Total Number of Observations - 41188
Total Number of Files - 1
Total Number of Features - 21
Base Format of the File - CSV File
Size of the Data - 5.699 MB

These are the categorical columns and what they represent:

- Job - Represents the job type of the customer which can be blue collar,technician,services,management,entrepreneur,self employed,housemaid,student etc
- Marital - Status of the customer which can be married, single, divorced or unknown
- Education - Level of education of the customer.That is: University degree,HighSchool,Professional Course etc
- Default - Whether the customer has a credit default or not.Can also be unknown
- Housing - If the customer has a housing loan or not.
- Loan - Whether the customer has a personal loan.
- Contact - Type of communication used to contact the customer i.e - Cellular,Telephone
- Month - Last contact month of the year
- day_of_week - Day of the week when the customer when last contacted
- poutcome - Outcome of the previous marketing campaign.It can be failure,success or nonexistent
- y(Has yes or no values which represent whethere the customer will purchase the term deposit or not)

Dataset Overview

These are the numerical columns and what they represent:¶

- Age - Age of the customer
- Duration - Duration of the last contact with the customer in seconds
- Campaign - Number of contacts performed during the campaign for the customer
- pdays - Number of days since the customer was last contacted in the previous campaign
- previous - Number of previous marketing campaigns in which the customer was contacted before the current campaign.Eg - 4561 customers were contacted during the first previous campaign and so on.
- emp.var.rate - Represents the employee variation rate. Whether the employment rate decreased or increased by the indicated percentage.A positive value indicates an increase in employment levels while negative values indicate a decrease in employment levels.
- cons.price.idx - Consumer Price Index.It's a measure of inflation where higher values indicate inflaction and lower values indicate deflation.
- cons.conf.idx - Consumer Confidence Index.Measures consumer's confidence about the state of the economy.Values closer to zero inidcate optimism in their financial situation and the economy's future.Values further from zero indicate pessimism in their financial institution
- euribor3m - Euribor 3-month rate.It represents the average interest rate at which European banks lend unsecured funds to one another for a three-month period.Higher values indicate higher short-term borrowing costs in the Eurozone hence higher market demand for funds and vice versa.
- nr.employed - Number of employees

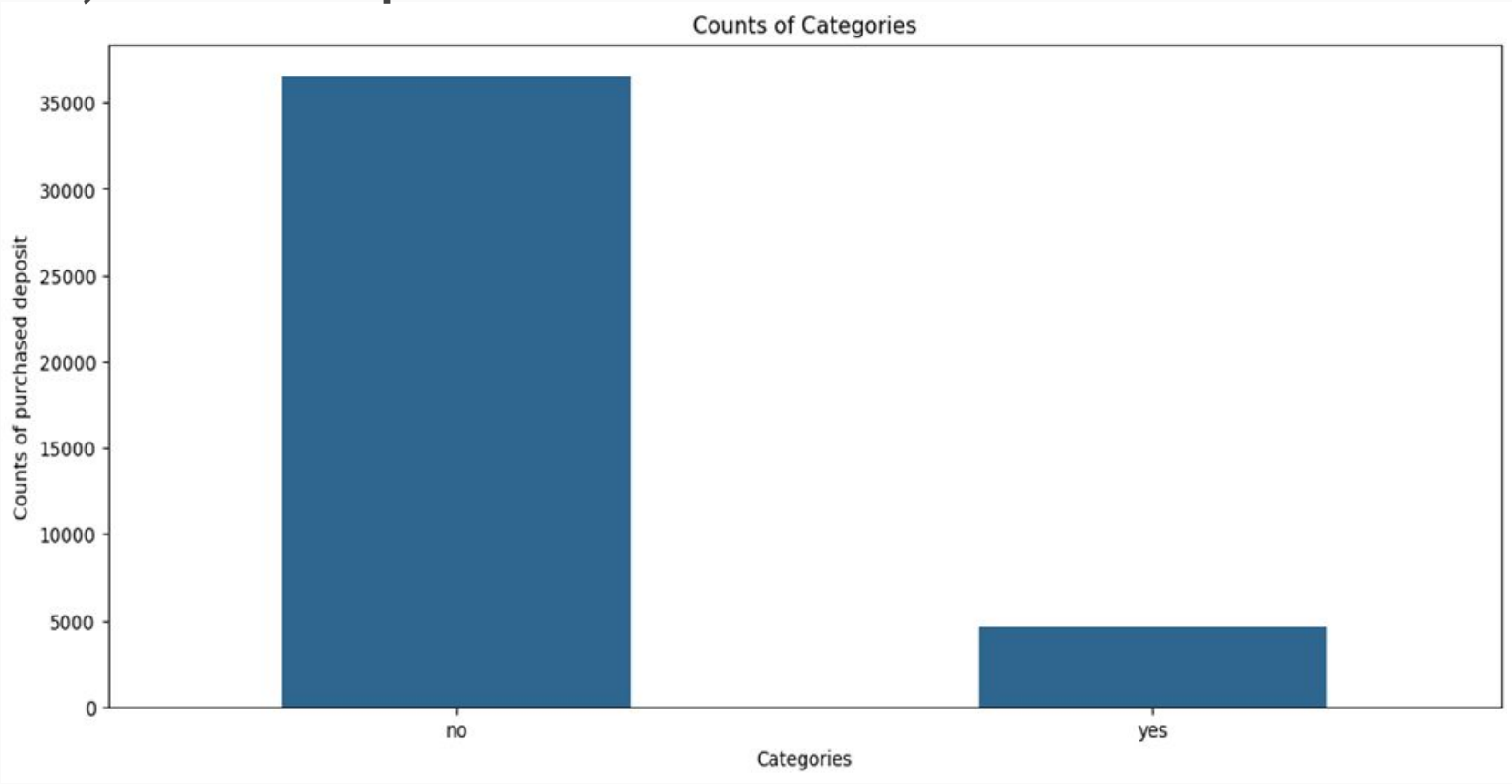
DATA CLEANING

I took these steps for data cleaning:

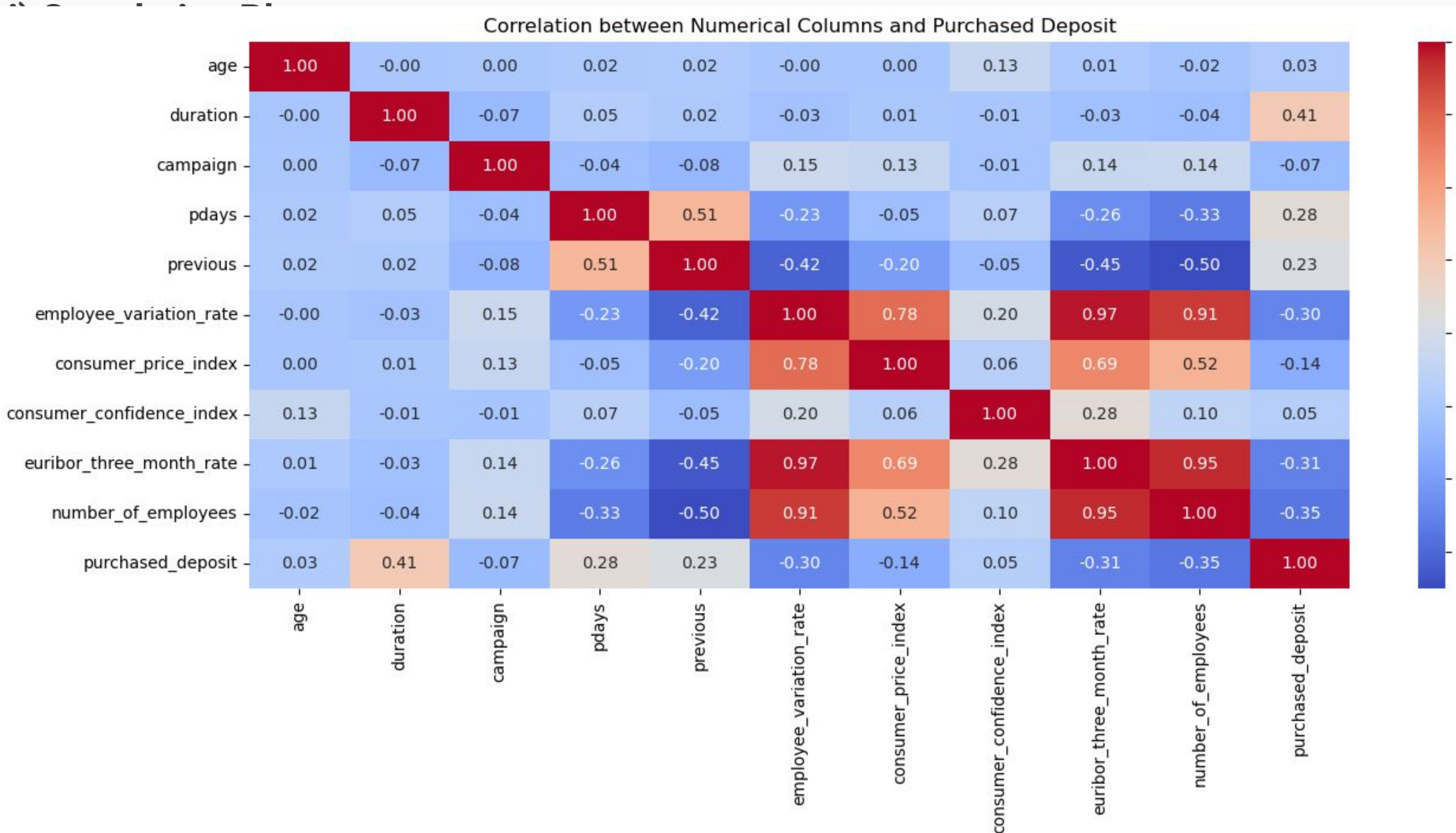
- Renaming columns
- Checking for outliers
- Checking for null values and duplicates
- Confirming if values of columns are in the correct format

EXPLORATORY DATA ANALYSIS

i) Purchased Deposit Counts



Most people did not purchase the term deposit


Correlation between Numerical Columns and Purchased Deposit
age 1.00 -0.00 0.00 0.02 0.02 -0.00 0.00 0.13 0.01 -0.02 0.03
duration -0.00 1.00 -0.07 0.05 0.02 -0.03 0.01 -0.01 -0.03 -0.04 0.41
campaign 0.00 -0.07 1.00 -0.04 -0.08 0.15 0.13 -0.01 0.14 0.14 -0.07
pdays 0.02 0.05 -0.04 1.00 0.51 -0.23 -0.05 0.07 -0.26 -0.33 0.28
previous 0.02 0.02 -0.08 0.51 1.00 -0.42 -0.20 -0.05 -0.45 -0.50 0.23
employee_variation_rate -0.00 -0.03 0.15 -0.23 -0.42 1.00 0.78 0.20 0.97 0.91 -0.30
consumer_price_index 0.00 0.01 0.13 -0.05 -0.20 0.78 1.00 0.06 0.69 0.52 -0.14
consumer_confidence_index 0.13 -0.01 -0.01 0.07 -0.05 0.20 0.06 1.00 0.28 0.10 0.05
euribor_three_month_rate 0.01 -0.03 0.14 -0.26 -0.45 0.97 0.69 0.28 1.00 0.95 -0.31
number_of_employees -0.02 -0.04 0.14 -0.33 -0.50 0.91 0.52 0.10 0.95 1.00 -0.35
purchased_deposit 0.03 0.41 -0.07 0.28 0.23 -0.30 -0.14 0.05 -0.31 -0.35 1.00
age
duration
campaign
pdays
previous
employee_variation_rate
consumer_price_index
consumer_confidence_index
euribor_three_month_rate
number_of_employees
purchased_deposit
1.0
0.8
0.6
0.4
0.2
0.0
−0
−0

Observations from the correlation plot

- Duration is moderately positively correlated to purchased deposit. It has a positive correlation of 0.41
- Pdays and previous are weakly positively related to purchased deposits. With a correlation of 0.28 and 0.23 respectively.
- Age has a very weak positive correlation of 0.03 to purchased deposit.
- The following columns are negatively correlated to purchased deposit:

  campaign, employee_variation_rate, consumer_price_index, euribor_three_month_rate, number_of_employees.

DATA PREPROCESSING

I performed the following data preprocessing steps on the data:

- One Hot Encoding
- Cyclic Encoding
- Dropping columns that are less efficient in our dataset.They have no effect on the target variable
- Normalization of the data
- Splitting the dataset to train and test
- Solving class imbalance

MODELLING

I have used the following models in my project:

1.Logistic Regression(BaseLine Model)

2.Random Forest(Ensemble Model)

3.XGBoost(Boosting Model)

The following table shows the comparative performance:

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Logistic Regression	90.36	65.50	38.52	48.51
Random Forest	90.51	62.42	48.92	54.85
XGBoost	91.27	66.45	52.42	58.61
XGBoost(After Tuning)	91.54	68.87	51.49	58.93

EVALUATION

I have used **F1-Score** as my Evaluation Metric for the following reasons

Accuracy - Measures overall accuracy

Precision - When False Positives matter the most

Recall - When False Negatives are more important

F1-Score - Best used when both precision and recall are important

For my project, Precision(False positives) means predicting that the customer will purchase the term deposit and they do not.This leads to waste of marketing resources which we do not want to be the case.

While Recall(False Negatives) means predicting that the customer will not purchase the term deposit and they actually do,leading to missed opportunity for revenue which we also do not want to be the case.

I used F1-Score to cater for both situations.

4.RECOMMENDATIONS

1.Develop marketing campaigns that highlight the benefits of term deposits, such as security, returns, and suitability for long-term goals by promotions or incentives eg slightly higher interests for higher customers. The bank can consider holding webinars and workshops to educate customers on the importance of term deposits.

2.Increase the reach and frequency of marketing campaigns, particularly focusing on customers who were not contacted previously. By considering using cellular to contact customers as it is yielding more results compared to telephone.

3.Focus marketing efforts during periods of high consumer confidence or low inflation (e.g., early in the year or after mid-year). The bank can monitor economic indicators like the Consumer Confidence Index and Euribor rates to plan campaigns strategically.

4.Train agents to have meaningful, informative conversations that address customer concerns within the call duration 0-500 seconds as most calls are within that range. They can provide training to call agents on how to effectively pitch term deposits.