Week 8: Data Understanding

Group Name: Data Visionaries

Name: Abdukhakimov Asatilla

Email: tremendous54@gmail.com

Country: Uzbekistan

Company: Data Glacier

Specialization: Data Science

Problem Description

ABC Bank wants to sell its term deposit product to customers. Before launching the product, the bank aims to develop a model to understand whether a particular customer will buy their product or not, based on the customer's past interaction with the bank or other financial institutions.

Business Understanding

ABC Bank aims to leverage a Machine Learning (ML) model to accurately identify customers who are most likely to subscribe to its term deposit product. By focusing marketing efforts on high-probability customers, such as through targeted telemarketing, SMS, or email campaigns, the bank can achieve substantial cost savings. This is made possible by reducing resource allocation to less likely prospects, thus optimizing time, effort, and budget.

The primary goal of developing this ML model is to enhance the bank's customer targeting strategy, ensuring that the marketing team directs its resources to the most promising leads. As a result, the bank will improve conversion rates while minimizing wasted marketing expenditures.

The model will be built in two distinct scenarios: one that incorporates the 'duration' feature and another that excludes it. While the 'duration' feature may improve predictive accuracy, it is not ideal for deployment due to its complexity and the potential lack of interpretability for business stakeholders. Furthermore, the 'duration' is typically only available after a customer interaction, making it an impractical basis for pre-call campaign strategies.

Therefore, model performance will be assessed and compared across both scenarios to strike the right balance between technical robustness and business feasibility. The focus will be on delivering a model that is not only effective but also practical, transparent, and aligned with the bank’s operational needs and marketing objectives.

Data Understanding

```
data = pd.read_csv('bank-full.csv')
```

```
data.head()
```

	age;"job";"marital";"education";"default";"balance";"housing";"loan";"contact";"day";"month";"duration";"campaign";"pdays";"previous";"poutcome";"y"
0	58;"management";"married";"tertiary";"no";2143...
1	44;"technician";"single";"secondary";"no";29;"...
2	33;"entrepreneur";"married";"secondary";"no";2...
3	47;"blue-collar";"married";"unknown";"no";1506...
4	33;"unknown";"single";"unknown";"no";1;"no";"n...

Our CSV file is separated by ; instead of ,, causing parsing errors where the entire row is read as a single column. Mixed delimiters, semicolons within data values, or inconsistent headers can further lead to misalignment and incorrect splitting. First, we should handle this problem.

Data Types

The dataset consists of both numerical and categorical data.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

Problems in the Data:

- Missing Values (NA): Some columns have "unknown" values (e.g., Job, Education), and pdays uses 999 to indicate no previous contact.
- Outliers: Columns like Age, Duration, and Campaign might have extreme values.
- Skewed Data: Columns like Duration and Campaign could have skewed distributions, meaning most values are clustered on one side.

Approaches to Handle Problems:

- Missing Values: Replace "unknown" with a separate category or impute missing values with the most frequent value. Treat 999 in pdays as missing or impute.
- Outliers: Use boxplots or Z-scores to identify and remove or cap extreme values.
- Skewed Data: Apply log transformations or square root to columns like Duration and Campaign to make them more normal.

Github Link:

https://github.com/asat94/Data-Glacier-Internship/tree/main/Week%208%20Data%20Understanding