**INFO 5307 – Project Part 3**

Banking & Finance (Market Trend Analysis Solution)

* Meghana Patibandla (11602295)

**System Design Project Report on Market Trend Analysis in Banking and Finance**

* I will be posting my analysis & Metrics in [**Github Link**](https://github.com/Meghana-Patibandla/INFO-5307--Knowledge-Management-Tools-and-Technologies-FALL-23)**.**
* The .ipynb file for the project :- [**Python File**](https://colab.research.google.com/drive/1joVo7CppWuaVUoimZtbCisoqFRYFXO_I?usp=sharing)
* The dataset link is gathered from [**here**](https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset)**.**
* The IDEI used to be Google Collab.
* The RoadMap I used to create this project: [**Monday**](https://view.monday.com/5368979012-89309f6f7e673d5124fd87e18115f538?r=use1)**.**

**Executive Summary**

**Objective:**

This project aims to enhance the effectiveness of marketing campaigns in the banking and finance sector by predicting customer responses to term deposit campaigns using the XGBoost machine learning classifier model.

**Introduction**

**Business Problem:**

The primary challenge for banks and financial institutions is optimizing marketing strategies to increase the conversion rates of term deposit subscriptions. The bank I chose is “Central Bank of India” where it mostly follows Traditional methods which often resulted in ineffective targeting and high campaign costs. In this, I will be addressing a solution for it.

**Project Scope:**

Leveraging data-driven insights to understand customer behavior and predict their response to marketing initiatives, thereby enabling more focused and efficient marketing strategies.

The project emphasizes the significance of feature engineering in creating meaningful predictors for customer responses. The XGBoost model is chosen for its ability to handle complex relationships within the data and provide interpretable results. I planned to divide the dataset into training and testing sets, implement techniques, and fine-tune hyperparameters to ensure model robustness. The expected outcomes would not only include improved campaign conversion rates but also a deeper understanding of customer behavior that could inform future marketing strategies.

**Architecture Diagram: -**

I used the Tool Lucid Chart to create this diagram. [Access here](https://lucid.app/lucidspark/80ba5728-0f2a-465f-a2f5-ca8e6bb5db98/edit?viewport_loc=-1114%2C-681%2C3138%2C1787%2C0_0&invitationId=inv_6a97a9f0-dbee-41d5-89a6-b7f076eccd4a) **A diagram of a business process

Description automatically generated**

**Methodology**

1. **Data Collection and Import:**

* **Dataset:** '[bank.csv](https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset)', encompassing customer demographic information and campaign interaction data.
* **Columns:** Key columns include age, job, marital status, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, and poutcome.

These columns provide a comprehensive view of the customer's demographic, financial status, and their interaction history with the bank's marketing campaigns, which are crucial for building a predictive model to analyze market trends in the banking and finance sector.

1. **Data Preprocessing:**

* **Dummy Variables Creation:** Transforming categorical variables such as 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', and 'poutcome' into dummy variables.
* **Data Cleaning:** Addressing missing values, normalizing numerical data like 'balance', 'duration', 'campaign', 'pdays', and 'previous'.

**Original Categorical Data:**

|  |  |  |
| --- | --- | --- |
| **CustomerID** | **Job** | **Marital Status** |
| 1 | admin. | married |
| 2 | technician | single |
| 3 | blue-collar | divorced |

**Transformed Data with Dummy Variables:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CustomerID | Job\_admin | Job\_technician | Job\_blue\_collar | Marital\_married | Marital\_single | Marital\_divorced |
| 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 1 |

In this:

* Each unique value in the 'Job' and 'Marital Status' columns is converted into a separate column.
* The presence of a category is marked by a 1, and its absence by a 0.

**3. Exploratory Data Analysis:**

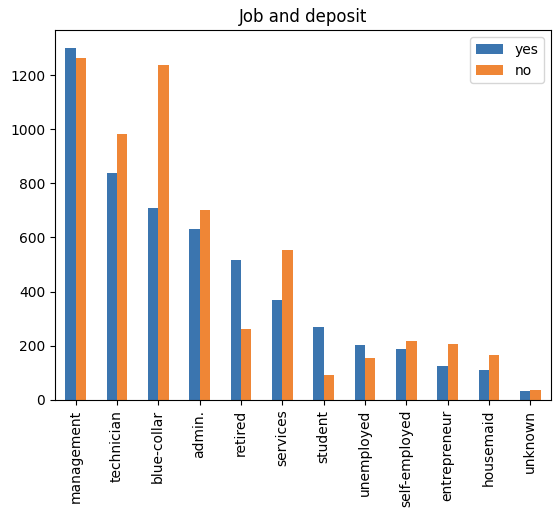
* **Categorical Data Analysis:** Examining the distribution of values in categorical columns and their correlation with the 'deposit' column (the target variable).  
    
  A screenshot of a graph

  Description automatically generated
* **Numerical Data Analysis:** Analyzing numerical columns like 'age', 'balance', and 'duration' to understand their influence on the campaign outcome.

**A group of blue and white bars

Description automatically generated**

We can see that numerical columns have outliers (especially 'pdays', 'campaign' and 'previous' columns).

**** **A graph of different colored bars

Description automatically generated** A graph of different colored bars

Description automatically generated

from the above analysis.

* Clients with employment classified as "blue-collar" or "services" are less likely to sign up for term deposits.
* Customers who are married are less likely to sign up for term deposits.
* Clients who use "cellular" contact are less likely to sign up for term deposits.

**A graph of a number of blue and orange bars

Description automatically generatedA graph with blue and orange bars

Description automatically generatedA graph with blue and orange bars

Description automatically generated**

From the above analysis we can conclude that:

* Term deposit subscribers often have higher balance and age values.
* Individuals who signed up for a term deposit often had fewer interactions during this campaign.

**4. Model Development (XGBoost Classifier Model):**

I used cleaned dataset for prediction of campaign outcome with help of machine learning classification models and used **XGBoost** Classifier which is one of the most common machine learning libraries for modelling.

**Data Preparation:**

* The cleaned and processed dataset was used, where categorical variables were transformed into dummy variables, and irrelevant columns were dropped or modified.
* The dataset was then divided into features (X) and the target variable (y), where 'y' was the likelihood of a customer subscribing to a term deposit.  
  A screenshot of a black and white list

  Description automatically generated

**Splitting the Dataset:**

* The dataset was split into training and testing sets with the 70 30 split. This is a standard practice in machine learning to evaluate the model's performance on unseen data.

A computer screen with text and images

Description automatically generated

**Training the XGBoost Model:**

* An XGBoost classifier model was initialized with specific parameters (like n\_estimators, learning\_rate, gamma, subsample, colsample\_bytree, max\_depth).
* The model was trained using the training set. XGBoost is known for its efficiency and effectiveness in classification tasks.  
    
  From the model Training, the main outcomes are:
  1. Clients who are older are more likely to sign up for a term deposit.
  2. Greater account balances increase the likelihood that a customer may sign up for a term deposit.
  3. Number of contacts with the customers really matters. Too many contacts with the customer could make him decline the offer.

**Feature Importance Analysis:**

* The XGBoost model provides a feature importance analysis, which was utilized to understand which features most significantly influence the prediction. This is critical for gaining insights into which aspects are most influential in a customer's decision to subscribe to a term deposit.

A graph with blue lines and black text

Description automatically generated

As we can see from the analysis of bar chart showing feature importance’s, the most important features are:

1. Customer's account balance,
2. Customer's age,
3. Number of contacts performed during this campaign and contact duration,
4. Number of contacts performed before this campaign.

**The main outcomes of the modelling are:**

1. Customers of greater age are more likely to subscribe for the term deposit.
2. Customers with greater account balance are more likely to subscribe for the term deposit.
3. Number of contacts with the customers really matters. Too many contacts with the customer could make him decline the offer.

**5. Model Evaluation and Optimization:**

* Performance Metrics: Using accuracy to evaluate the model’s predictive capabilities.
* Feature Importance Analysis: Identifying which features most significantly affect the model's predictions.

**More specific Recommendation**  
We also focused on Finding out account balance, which marketing campaign should focus on

A graph with blue lines

Description automatically generated A graph with blue line

Description automatically generated

From the above analysis,

* We can say that marketing campaigns should concentrate on customers with account balance greater than 1490$.
* Average subscrition rate tends to be higher for customers below 31 years old or above 56 years old.

**Results**

**Model Performance:**

* The model demonstrated high accuracy and precision in predicting customer responses, XGB accuracy score for training is 91 % and with test data is 85 %

A computer screen shot of a program code

Description automatically generated

**Conclusion:**

This analysis provides a strategic framework for targeted marketing, significantly enhancing campaign efficiency. The results provide a clear roadmap for future marketing efforts, empowering financial institutions (banks) to make informed decisions and stay adaptive in a dynamic market landscape.

**Strategic Recommendations for Future Work:**

1. **Account Balance Influence:**

* From the Analysis Customers with more than $1490 in their account are more likely to subscribe to a term deposit. Focus on reaching out to these customers in future campaigns.

1. **Age Influence:**

* The age of the customer matters for campaign success. Concentrate on customers below 30 and above 50 years old in future campaigns.

1. **Number of Contacts:**

* Limit the number of contacts with a customer during the campaign to 4 or fewer, as it significantly impacts the campaign outcome.