Bank Subscription Prediction

Ansh Lulla | 22070126013 | AIML A1

Problem Statement - Need analysis of the application

At the core, the retail banks operate on a simple principle: it's that it borrows money at a low interest rate and lends it out at a higher interest rate. The difference between the interests earned from loans and the interests that are paid on deposits is called the **Net Interest Margin (NIM)**. This margin acts as a source of profitability for the banks.

Term deposits are important especially because they ensure a stable and a predictable source of capital which is essential for bank's financial decisions and planning alongside risk assessment. Previously acquisition of these deposits depended on high-cost marketing which were inefficient and time consuming (also referred to as the shotgun approach). Machine Learning on the other hand aims to provide rather a rifle-approach where it would analyze the data, identify clients which are most probable to subscribing to the bank and thus makes the process smarter and more effective. This modern AI-based approach can ensure maximization of ROI in marketing, improve operational efficiency and improve user experience.

Impact Overview Statement

This application can potentially serve several benefits both on the operational-side and revenue-side.

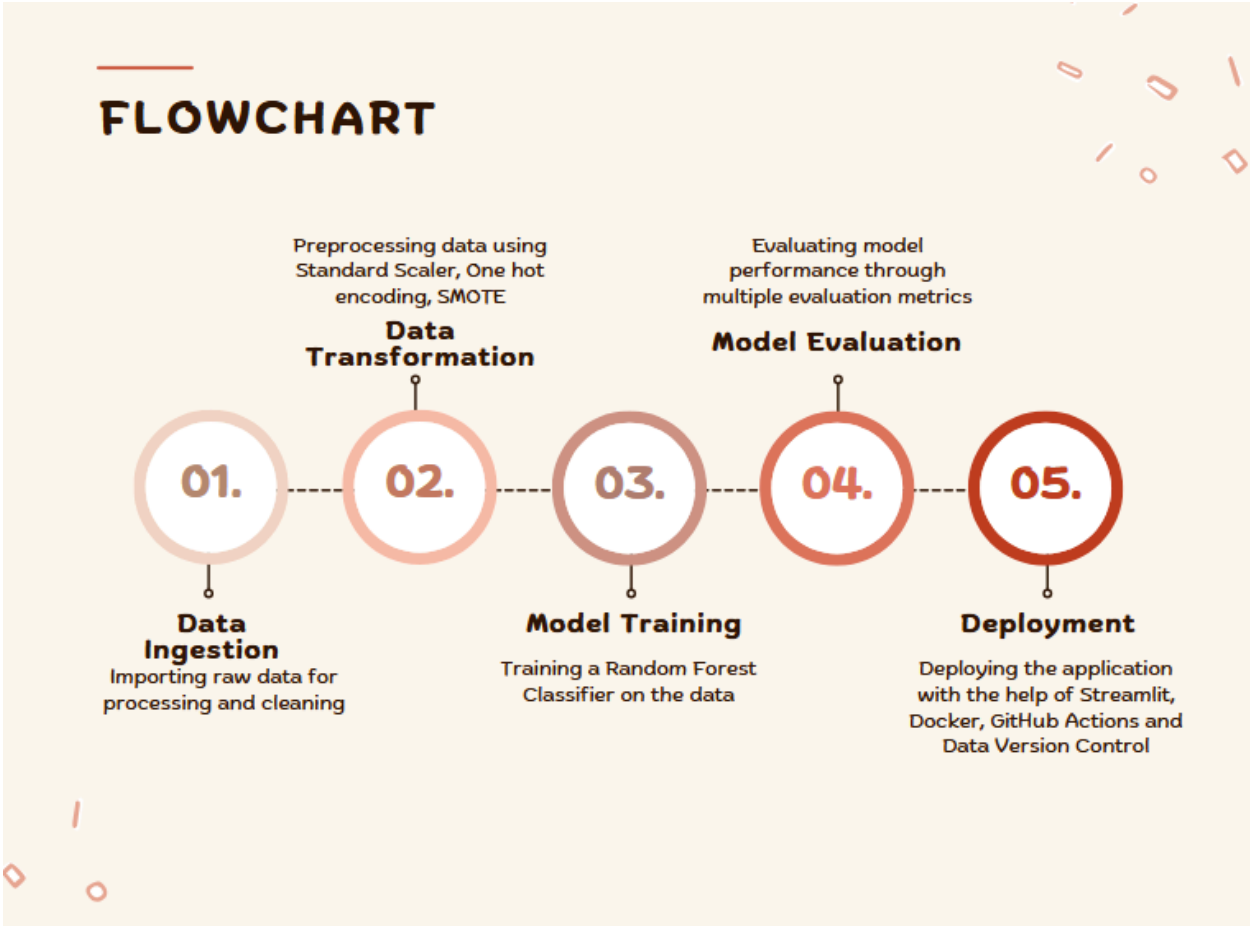
1. **Optimized marketing campaigns:** This approach allows the bank to focus its time and efforts only on the targeted and potential customers who would be willing to subscribe, which would enhance the productivity and outcome of the marketing campaigns, and hence lead to potentially higher returns.
2. **Improved Customer Relations:** Understanding the behaviour and history of the customers, the banks can tailor its products and services in a better way,

which are more customer-centric as well, resulting in enhanced customer experience and better relations of the bank with its customers.

3. **Cost Reduction:** Targeting the right customers helps cut down the costs related to marketing campaigns and similar call center tasks.
4. **Increase in Revenue:** A higher subscription rate would lead to higher increases in the revenue of banks which could potentially act as a primary source of funds and investments made by the banks.

Architecture

This project implements an end-to-end ML Pipeline in order to predict whether a customer will subscribe to a term deposit. This methodology ensures reproducibility, modularity and production-ready. The steps are described below:



1. **Data Ingestion:** The raw dataset is read into the system and a clean copy is persisted in the directory for ease of access and convenience. This avoids the pipeline to re-run and re-load the data at every step hence making the pipeline more efficient.
2. **Data Transformation and Preprocessing:** The dataset consists of numerical and categorical variables so it's important to preprocess the data carefully.
    a. Categorical Variables are preprocessed using One Hot Encoding and the Numerical Variables are preprocessed using Standard Scaler to bring variables like age and salary to the same range so that the model gives equal importance to all the variables.
    b. The model might get skewed toward variable like salary due to its high numbers so it's important to scale the variables to lie in the same range for same levels of comparisons.
    c. The dataset is also split into train and test splits for model training and model testing respectively (following a 80-20 split).
    d. The dataset inherently also had imbalanced classes (a large number of "not subscribed) and a smaller number of "subscribed" customers) this would lead to misclassification and overfitting of the model if not treated properly. To treat this, SMOTE (Synthetic Minority Oversampling Technique) was used which helps proportionate the training samples for "not subscribed" and "subscribed" during training so that the model does not get skewed towards predicting a single class only.
3. **Model Training:** A classification model (Random Forest Classifier) was trained on the training dataset with hyperparameters tuned for balanced and faster training. The trained model is saved as a pickle file to be re-used at the time of deployment
4. **Model Evaluation:** Evaluating the performance of the model is equally important to training of the model so multiple evaluation metrics have been used to evaluate the model's performance. The evaluation metrics used:
    a. **Accuracy:** Total number of predictions which were truly positive or truly negative out of all the predictions made.

b. **Precision:** Out of all the predictions which were positive, how many were truly positive. This is important to rule out the false positives as we should not focus on customers not likely to subscribe.

c. **Recall:** Out of all the actual positives, how many of them were predicted positives. This helps rule out the false negatives in the predictions.

d. **F1-Score:** The harmonic mean of Precision and Recall. Provides a balance of both, precision and recall.

The metrics have been saved as a JSON object for future references.

5. **Data Versioning and Pipeline Orchestration through DVC:** Data Version Control (DVC) is used to version data, artifacts, and the pipeline itself. All the stages are automated in a yaml file which ensures reproducibility of the system and pipeline automation.
6. **Deployment:** The project is deployed via streamlit and is ready-to-use on the streamlit community cloud. User can either upload an entire excel file (which follows the schema of the data used to train the model) or manually input the values of the attributes required to predict the outcome of a subscription.
7. **Containerization via Docker:** The application is containerized via Docker for environment consistency and versatility. It helps run the application anywhere and resolve any dependency issues related to migrations.
8. **CI/CD Pipeline:** Github Actions have been used to automate the entire pipeline which is triggered whenver a new push or pull request occurs in github. This ensures consistency of performance even if the code or the data changes.

Tech Stack Used

1. Programming Language: Python
2. Libraries and Frameworks: pandas, numpy, scikit-learn, dvc, dotenv, pydantic, gdown, imblearn
3. Deployment: Docker, GitHub Actions, Streamlit
4. OS - Ubuntu Linux

Model Performance

The below JSON object is metrics.json, saved after evaluating model's performance.

```
# Metrics.json
{
   "accuracy": 0.8914077186774301,
   "precision": 0.5416348357524828,
   "recall": 0.6498625114573785,
   "f1score": 0.5908333333333333
}
```

Project Usage

1. **Streamlit App:**

   Live Deployment at: Streamlit Deployment

2. **Git Steps:**

   ```
   git clone https://github.com/Anshlulla/AIBF-Bank-Subscription-Prediction
   cd AIBF-Bank-Subscription-Prediction
   pip install -r requirements.txt
   pip install dvc
   dvc repro
   streamlit run app.py
   ```

3. **Docker Support:**

   ```
   docker build -t anshlulla/aibf-project:latest .
   docker run --rm -p 8501:8501 aibf-project:latest
   ```

Conclusion

This Application demonstrates how Machine Learning can be used to reduce manual labour and increase the efficiency of bank's marketing campaigns so that the bank knows beforehand which customers to target in order to get the

customers to subscribe to the term deposit, thus helping the bank make smarter and more profitable financial decisions and investments. By leveraging versioning for both, the code and the data helps optimize the workflow, induce best practices for model training and performances.

GitHub Repository Link

GitHub Repository