Group Name: GIG group

Name: Rupert Tawiah-Quashie

Email: rupertquash@gmail.com

Country: USA

College: Hampshire College

Specialization: Data Science

Problem Description:

- ABC Bank currently sells term deposits (fixed-term savings accounts) through generalized marketing campaigns via channels like telemarketing, email, etc.
- They want to improve the efficiency of marketing by targeting customers more likely to subscribe to the term deposit product.
- The bank has data on ~45,000 customers with details on demographics, account history, previous marketing contacts, economic indicators, and most importantly the label of whether the customer subscribed to a term deposit in the past campaign.
- The goal is to build a predictive model using this data to estimate the probability that each customer will subscribe to a term deposit.

Business Understanding:

- The key business objective is to optimize marketing resource allocation and costs by focusing only on the highest potential customers predicted by the model.
- Current response rates for term deposit campaigns are estimated around 10-15%. If the model can better target marketing, this rate can hopefully be improved significantly.
- Increased term deposit subscriptions will provide a greater capital base for ABC Bank to expand lending activities and revenue opportunities.
- The model will need to integrate with existing marketing campaign systems and data workflows. Inputs will need to be extracted and predictions scored on a regular cadence.
- Model accuracy metrics like ROC AUC, precision, recall will be important. But business is most focused on directly improving response rate % among targeted subset.
- Data imbalances in the training set should be handled to ensure predictive power across customer segments. The full data should be utilized.
- Duration feature needs to be excluded given business constraints around explainability. Other useful features like age, job type, previous contacts should be leveraged.
- Model interpretability is also important for building business understanding of driving factors.

Project lifecycle along with deadline

Week 1 (Nov 19-25):

- Review problem statement and business goals in depth
- Clarify objectives and metrics for model success with stakeholders
- Import data, assess quality, check for missing values and outliers
- Explore distributions of features, correlations between features
- Identify promising features for predicting term deposit subscription
- Document initial data understanding and EDA findings

Week 2 (Nov 26-Dec 2):

- Clean data (handle missing values, remove unnecessary/duplicate features etc.)
- Engineer new features as needed based on insights from EDA
- Split data into train/validation/test sets
- Try logistic regression, random forest, XGBoost models as benchmarks
- Tune hyperparameters using cross-validation on train set
- Evaluate models on validation set, compare to baselines
- Handle class imbalance with techniques like SMOTE if needed
- Identify best performing modeling approach so far

Week 3 (Dec 3-9):

- Refine data processing pipeline based on insights from initial models
- Generate new feature combinations/transformations as needed
- Run multiple iterations of models, tuning hyperparameters
- Begin experimenting with model ensembling/stacking
- Continue evaluating on validation set and compare to baselines
- Monitor for overfitting, adjust regularization as needed
- Select 1-2 best performing models/ensembles

Week 4 (Dec 10-16):

- Finalize model training on full dataset
- Document model performance on test set with accuracy, AUC, etc.
- Convert model performance metrics to projected business value
- Prepare model interpretations and explanations for stakeholders
- Develop deployment plan and requirements for production
- Create presentation to explain approach, results, recommendations

Data Intake Report

Name: Bank Marketing

Report date: November 18, 2023

Internship Batch: LISUM26

Version: 1.0

Data intake by: Rupert Tawiah-Quashie

Data intake reviewer:

Data storage location:

Tabular data

bank-additional-full.csv

Total number of observations	41188
Total number of files	1
Total number of features	20
Base format of the file	.csv
Size of the data	5.8 MB

bank-additional.csv

Total number of observations	4119
Total number of files	1
Total number of features	20
Base format of the file	.csv
Size of the data	584 KB

bank-full.csv

Total number of observations	45211
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	4.6 MB

bank.csv

Total number of observations	4521
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	461 KB