Title: Customer Bank Account Opening Propensity Modeling

Introduction:

The objective of the project is two folds.

1. To Help marketing Executives at a large bank understand characteristics/ best predictors for bank product purchase.
2. To build a Predictive model to score potential customers.

Methodology:

Data collection:

The Bank marketing data used for our project was downloaded from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Exploratory Analysis: EDA

The training dataset consisted of ~ 42,000 records and 21 features and the test data consisted of 4100 records. Exploratory Data analysis was done primarily using Univariate and Bivariate plots and secondarily Tabular summaries and with other statistical techniques, where appropriate(e.g correlation, variance Check). It helped us identify several data issues present in the Data and to ensure the quality of the data i.e

a. Missing Values,
b. Feature Scale Issues
c. Data Distribution Issues
b. Outliers
d. Feature Values/ Frequency Validity
e. Correlated variables

It also helped us identify transformations that needed to be done on the features i.e

a. Categorical Encoding (Binary / Dummy Features)
b. Feature Transformation:
c. Eliminate Features

In addition to the visual observational analysis, we also use statistical tests during our EDA to easily identify and flag issues i.e highly correlated features and low variance features.

Reproducibility:

To ensure reproducibility, we have used the random seed to ensure that the randomization is the same across any number of runs. In addition, we have provided the code in a jupyter notebook alongside the Data, for anyone to rerun them and reproduce the results.

Preprocessing:

The Issues identified in the EDA led to the following preprocessing to

- Ensure Data Quality
- Ensure Modeling Data adheres to the upcoming ML model Data assumptions (e.g scaling- LR, no missing values- Sklearn-Trees, no correlated variables e.t.c)

It also helps the upcoming statistical model with better convergence and better predictive modeling power. The following preprocessing was done and applied to the data.

1. **Missing Values**:

1.1. Remove Records with missing values on features: (Marital - 0.19% missing). These records were removed because the size was small enough to be safely removed without losing too much Training data. It is also possible for these missing values to be part of Data collection or other processes, given their small size.

1.2. Fill: Out of mean, median, mode, interpolation, and proportionate fills available, we choose proportionate fill. And because these records were in moderate size and treating them as the separate feature value was not justified. The following features were proportionately filled (housing- 2.4% missing, loan - 2.4% missing, education - 4.2%, job- 0.8%)

1.3. Keep as a separate class: We treated some missing values as a separate class because the missing value size was not significant enough to eliminate the feature completely but was significant enough to be randomly filled. It is also possible that missing values by themselves represent some information e.g (default - 20% Missing, if is self-reported then it's possible that, there could have been a high correlation between missing(unreported) or defaulted at some point in time).

2. **Inadequate Feature Variable Frequency**: Some values of the feature columns were too few to be used reliably or make any inference. Hence such variables i.e Education-Illeterate (0.04%) and default- yes (0.01%) were removed.

3. **Categorical Encoding**: Some Models i.e Regression, NN, SVM cannot handle categorical values and need to be converted into Dummy variables or Binary Encoded Variables. The following transformations were done:

3.1. Dummy variables: If a feature has >2 categories, they were transformed into dummy variables. (month, education, job, outcome, day_of_week, marital, default)

3.2. Binary Encoding: Features with only two categories i.e yes/no were converted into binary variables (i.e loan=yes/ no -> has_loan=0/1) (housing, loan, contact).

3.2.1. Pre-transformation was necessary on pdays to binary encode. Because > 95% data was client_previously_not_contacted and only 5% continuous[1 to 40]. It was transformed into a was_previously_contacted binary feature variable.

4. **Scaling**: Having numerical features on different scales i.e 5000 vs 0-10 creates problems for models. They create slower non-convergence at best to worst performing model at worst. Hence we scaled the following features.

4.1. Nr.employed[5000 range values], cons.conf.IDX, cons.price.IDX, euribor3m, age

Of diff scaling methods available ie min-max scaling, standardization, normalization, percentile-scaling we used percentile-based scaling because it preserves outliers and is less outlier sensitive for non-outlier values.

Scaling makes prediction good. But because it increases model explainability/comprehension challenge later, we kept both scaled and original columns and later during modeling and will experiment using both feature variants alternatively. If there is no significant loss in prediction power, we will favor non-scaled columns for better and easier explainability, as explainability is also one of our objectives.

5. **Skewed Data Distribution**: Skewed and non-zero mean-centered data causes challenges for models and their convergence. Of log transform, squaring, and sq/cube rooting available, we did log transformation for highly skewed variables i.e (no_of_campaigns, previous, and age). Again as explained above, we keep both and experiment them out in favor of explainability/ no significant predictive power loss.

6. **Outliers**: We detected outlier features using the Tukey method (previous, campaign, age, cons.conf. IDX). Of clipping, Replacement(mean, median, mode, model-interpreted), transformation(log, binning) and Keep available, we choose to keep it because our model by itself is an outlier detection model (11%- yes class). Hence these outlier values are more likely to be signals for outlier detection rather than noise, hence we kept them for this iteration. In the future, we can test different outlier resolution strategies and test models.

7. **Binning**: Occasionally we may get better and significant model improvements with binning (i.e transforming numerical variables to categorize). We tested with/ without binning for age category in later Logistic Regression(LR) Modeling. Because LR produced more significant p-values for age_bins, we in a later stage chose them for our final model.

8. **Correlation**: Correlation creates problems for models. E.g For LR, they produce a numerically unstable model and can impact convergence. Hence we removed highly correlated variables (emp.var.rate & euribor3m)

9. **Variance**: Features with very low variance add very little information/ value to models. We checked and observed None for our current case.

10. **Multi-Collinearity**: ML Models i.e LR have no multi-collinearity data assumption. It leads to a numerically unstable model and for Lr in particular creates coefficient computation challenges. We removed multi-collinearity introduced with dummy variables for removing a reference variable i.e Remove- job_unemployed, month_dec, marital_single, day_of_wrrk_friday. We tested using different reference variables & observed their impact on p-values to make the final reference variable selection.

Modeling:

DataSets:

We did a stratified split of the Train Data into 70% Train Data and 30% Validation Data. The final Test set was used for final Model Validation purposes only, so as not to get an independent evaluation. Often with parameter tuning, we may accidentally fit the model to validation data. Hence independent held-out test set is preferred and kept to prevent overfitting.
Imbalanced Data: We also created synthetic data for training data (not validation/ test data)'s minority class using SMOTE to create a balanced Dataset. Synthetic Balanced / Unbalanced Data both were tested and experimented with during the later modeling stage.

Evaluation Metrics:

Because we had imbalanced data (yes-11% vs no-89%), we carefully choose metrics that are able to handle them well i.e recall, precision, f1-score,auc_roc, R-squared, AIC. Amongst them, we focussed more on recall, because we wanted to maximize capturing no of people who have opened accounts, at an acceptable cost (precision).

Model Building / Validation:

We started by building a benchmark model (Tree model) and then into building more complex explainable models (Logistic Regression) while benchmarking them against the original Benchmark model. More complex but unexplainable models i.e (Deep Neural Networks) although more powerful and have higher predictive power were excluded from our current scope because of a lack of explainability.

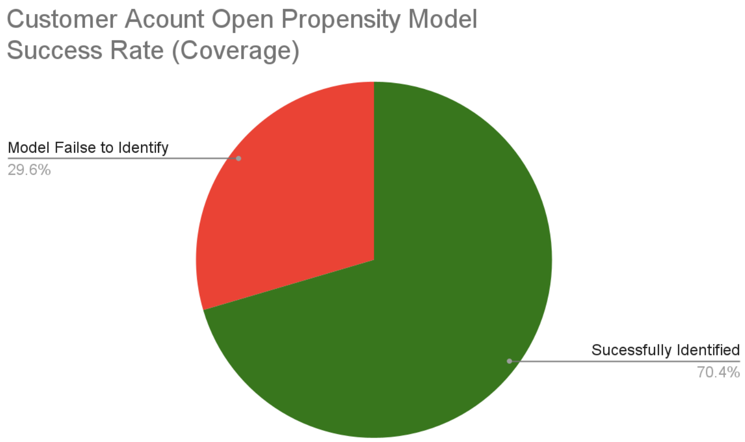
During modeling, we tested models with different data_variants(i.e scaled features vs non-scaled features, skew_fixed variables vs original variables.) and parameters(including regularisation parameters.)

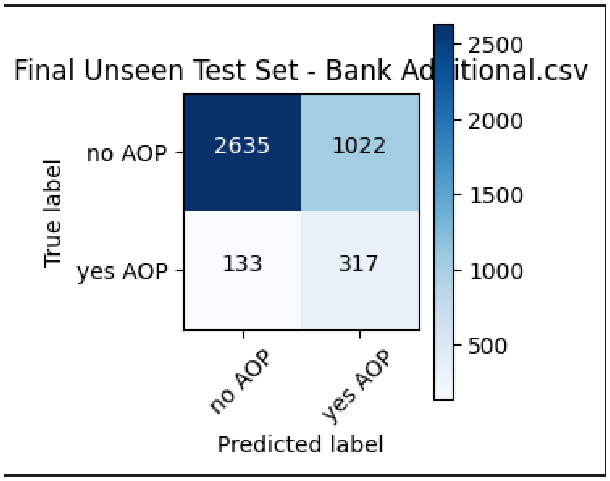
We used Sklearn-Logistic Regression for its powerful parameter customization and tuning capability. And later followed up with StatsModel-LR because they provide more model and coefficient statistical information primarily coefficients p-value & marginal effect.

Results:

The following result were obtained for our models.

model_name	**recall**	**precision**	**accuracy**	**balanced_accuracy**	**f1**	**roc_auc**
Simple Tree Unbalanced	0.35	0.31	0.84	0.62	0.33	0.62
Simple Tree balanced	0.32	0.29	0.84	0.61	0.3	0.61
Tree-Fix-MissVal-CatEncode	0.34	0.31	0.84	0.62	0.33	0.63
Tree-Fix-MissVal-CatEncode-Skew	0.34	0.31	0.84	0.62	0.32	0.62
Tree-Fix-MissVal-CatEncode-Skew-Scale	0.34	0.31	0.84	0.62	0.32	0.62
Tree-Fix-MissVal-CatEncode-Skew-Scale-Corr	0.33	0.3	0.84	0.62	0.32	0.62
LR-Fix-MissVal-CatEncode-C_0.0001	0.71	0.25	0.72	0.72	0.37	0.76
LR-Fix-MissVal-CatEncode-Corr-C_0.0001	0.59	0.34	0.82	0.72	0.43	0.76
LR-Fix-MissVal-CatEncode-Skew-C_0.0001	0.71	0.24	0.72	0.72	0.36	0.76
LR-Fix-MissVal-CatEncode-Skew-Scale-C_0.0001	0.71	0.24	0.72	0.72	0.36	0.77
LR-Fix-MissVal-CatEncode-Skew-Scale-Corr-C_0.0001	0.59	0.3	0.8	0.71	0.4	0.76
LR-SMOTE-Fix-MissVal-CatEncode-Skew-Scale-Corr-C_0.0001	0.66	0.3				
Final Selected Model : (Test Set) **LR-Fix-MissVal-CatEncode-Skew-Scale-Corr-C_0.0001**	**0.7**	**0.29**				



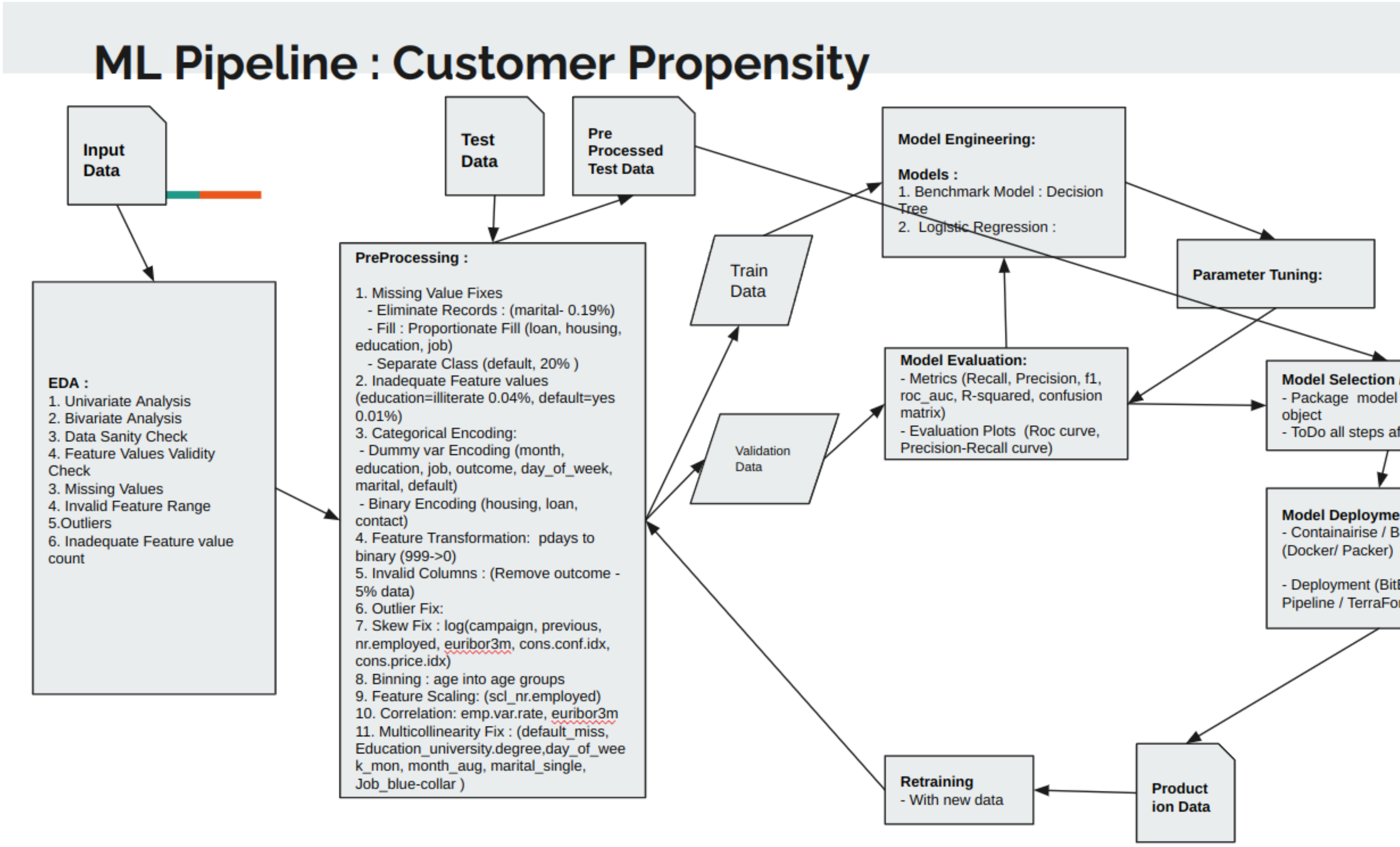


Model Selection:

We finally selected the LR model with R- squared value of 0.21.

Model Pipeline:

Please Refer to ML Model Pipeline for more detailed / better quality architecture.



ML Pipeline : Customer Propensity

Input Data

EDA :
1. Univariate Analysis
2. Bivariate Analysis
3. Data Sanity Check
4. Feature Values Validity Check
3. Missing Values
4. Invalid Feature Range
5.Outliers
6. Inadequate Feature value count

Test Data

Pre Processed Test Data

PreProcessing :

1. Missing Value Fixes
 - Eliminate Records : (marital- 0.19%)
 - Fill : Proportionate Fill (loan, housing, education, job)
 - Separate Class (default, 20%)
2. Inadequate Feature values (education=illiterate 0.04%, default=yes 0.01%)
3. Categorical Encoding:
 - Dummy var Encoding (month, education, job, outcome, day_of_week, marital, default)
 - Binary Encoding (housing, loan, contact)
4. Feature Transformation: pdays to binary (999->0)
5. Invalid Columns : (Remove outcome - 5% data)
6. Outlier Fix:
7. Skew Fix : log(campaign, previous, nr.employed, euribor3m, cons.conf.idx, cons.price.idx)
8. Binning : age into age groups
9. Feature Scaling: (scl_nr.employed)
10. Correlation: emp.var.rate, euribor3m
11. Multicollinearity Fix : (default_miss, Education_university.degree,day_of_week_mon, month_aug, marital_single, Job_blue-collar)

Train Data

Validation Data

Model Engineering:

Models :
1. Benchmark Model : Decision Tree
2. Logistic Regression :

Parameter Tuning:

Model Evaluation:
- Metrics (Recall, Precision, f1, roc_auc, R-squared, confusion matrix)
- Evaluation Plots (Roc curve, Precision-Recall curve)

Model Selection
- Package model object
- ToDo all steps af

Model Deployme
- Containairise / B
(Docker/ Packer)

- Deployment (Bit
Pipeline / TerraFo

Retraining
- With new data

Production Data

Discussion:

Our final model has an unscaled version of feature variables except for scl_nr_employed because scaling of Nr_employee had a significant positive impact on LR performance (R2 from 0.17 to 0.21). It is also reasonable because it has extremely high-scale data in a range of ~5000 compared to other features [~0-1]. We favored unscaled versions of other feature variables because they had no significant prediction power loss and helped us with better explainability.

Conclusion:

With our final model, the Bank will be able to successfully capture 70% (Recall) of customers who will open accounts. The cost to the bank will be that they will have to contact 3 more people for every successful conversion. I.e Our model has an accuracy of 29% (Precision) on the unseen test set. And Out of the 4 people our model recommends contacting, 1 will lead to an account opening and the rest 3 will be False Hits.

Based on our model, the top 5 positive and negative factors that lead to account opens were

1. **Top 5 Positive Features**

- Previously contacted: People contacted in previous campaigns have 5.8 odds of opening a bank account compared to not contacted ones.
- Months: People are more likely to open an account in March(odds: 2.7), June (odds: 1.5), and July(odds: 1.4), Dec compared to August(lowest rate month)
- Contact: People contacted in cellulars have 1.6 odds of opening a bank compared to people contacted in the telephone.
- Day of the week: People are more likely to open accounts in Tuesday(odds: 1.3), Wednesday(odds: 1.4), Thursday(odds: 1.3), and Friday(odds: 1.2) compared to Monday.
- Occupation: Students(odds: 1.4), retired people(odds: 1.3), or people working in admin(odds: 1.2) are more likely to open accounts compared to blue-collar workers

2. Top 5 Negative Features

- Number of employees: 1 unit increase in a Scaled number of employees decreases opening account propensity by 75%.(5000 scaled to 0)
- Months: People are less likely to open an account in May(odds: 0.6), Sep(odds: 0.7), and Nov(odds:0.8) compared to August.
- Age: People of age group 35-45(odds : 0.75), 45-55(odds : 0.8), 25-35(odds : 0.85) have lower odds of opening account.
- Default: People who have defaulted have 0.78 odds of opening an account compared to
- Education: People with basic_4y(odds: 0.9), and basic_9y(odds: 0.9) have lower odds of account opening compared to university_degree.

Future Work:

Given the limited time frame, not all potential action items could be pursued. However, we have identified and listed them for future work

- Include Interaction Terms in Logistic Regression
- Analyse model parameters against the EDA bivariate plots and discuss/ validate.
- Test more different Preprocessing Strategies and their impact on models. (e.g outlier removal, clipping)
- More variable slicing/ segmentation for significant p-values
- More Complex Modeling (Deep Neural Network) for more predictive power.