BANKING INSURANCE PRODUCT – PHASE 2

ORANGE HOMEWORK TEAM 8

Brendan Bammer
Karthick Krishna Balaji
Elisa Marchione
Nish Torane
Yvonne Tsai

November 18, 2022

Table of Contents

BANKING INSURANCE PRODUCT – PHASE 2

Overview

The Customer Services and New Products Department at Commercial Banking Corporation (hereafter the "Department") tasked Orange Homework Team 8 (hereafter the "analysts") with predicting which customers will purchase a variable annuity product. The Department provided 37 relevant variables related to customers' personal information, existing accounts, and service interactions to aid the prediction.

The analysts modeled two tree-based algorithms in Python: one using Random Forest algorithm (hereafter "RF") and the second using Extreme Gradient Boosting (hereafter "XGBoost"). The analysts trained these algorithms with parameter tuning to enhance their predictive power. This resulted in an Area Under Receiver Operating Characteristic curve (ROC-AUC) of the RF model being 0.88 and the ROC-AUC for XGBoost model being 0.93. It is important to note the tradeoff between the predictive power of these models with their lack of interpretability.

From the developed models, the analysts have three recommendations for the Department:

- Use the XGBoost Grid Search model to predict the purchase of the variable annuity product.
- Look into branches 14, 15, 18, and 19 because these branches do not have observations for Investment Account and Credit Card variables. The analysts recommend investigating why these branches are not offering these products.
- Evaluate model effectiveness based on an unseen dataset.

Methodology and Analysis

Data Used

The data contains information about customers who have bought a variable annuity insurance product and those who have not. The training dataset has 8,495 observations and 38 variables, including the target variable. As per the client's request, variables with more than ten unique values were converted to continuous, excluding the Branch variable. The analysts also dealt with missing values using median imputation for continuous variables and mode imputation for categorical variables. Finally, before creating a model, the analysts converted the Branch variable into dummy variables and dropped Total Amount for Point of Sale Interactions and Money Market Balance due to multi-collinearity. The analysts created a random distribution variable to use as a reference while performing the variable selection since this variable has no meaning for this use case. The final dataset has 8,495 observations and 54 variables, including the target variable.

Random Forest Classifier

After preprocessing the data, the analysts began modeling with a Random Forest classification model. The analysts used two cross-validation techniques to identify the best parameters to run the model. The analyst first performed Randomized Search cross-validation to determine the hyperparameters which were used as a reference point for Grid Search cross-validation. The analysts used Out of Bag Error (OOB) to ensure no overfitting issue. The analysts performed hyperparameter tuning on seven parameters on both the cross-validation techniques. These seven parameters are in **Table 1** on page 2.

Table 1: Hyperparameter Tuning Output on Random Forest

Parameter	Randomized Search CV	Grid Search CV
Number of trees in Random Forest	200	1000
Number of features for consideration at every split	Square Root of Total Number of Features	Square Root of Total Number of Features
Maximum number of levels allowed for each tree	72	200
Minimum sample for split	10	15
Minimum sample for leaf	1	3
Bootstrap	True	True
Out of Bag Samples to estimate generalization error	True	True

Table 1 shows the different values for the parameters that were obtained by running cross-validations. The analysts ran the model using these parameters, and the resulting OOB error and ROC-AUC score are in **Table 2**.

Table 2: ROC-AUC Score and Out-Of-Bag Error Score on Random Forest

Cross Validation Technique	Out-Of-Bag Error Score	ROC-AUC Score
Randomized Search CV	0.73	0.94
Grid Search CV	0.74	0.88

From **Table 2,** the analysts noticed that Randomized Search cross-validation has performed better than Grid Search cross-validation. The variable selection and top 10 variables ranked by Gain impurity for this model are in **Figures 3** and **4** in the **Appendix**.

Extreme Gradient Boosting Classifier (XGBoost)

The second model the analysts created was the XGBoost model. The RF model uses a bagging technique, whereas XGBoost uses a boosting technique for prediction. Like RF classifier, the analysts first performed Randomized Search cross-validation followed by Grid Search cross-validation to handle overfitting the data. The analysts performed hyperparameter tuning on seven parameters on both cross-validation techniques. **Table 3** on page 3 shows the final output with the tuned hyperparameters from the cross-validation techniques.

The analysts performed variable selection and identified that 24 variables were below the random variable created, as seen in **Figure 2** in the **Appendix**. The variable selection and top 10 variables ranked by Gain impurity for this model are in **Figures 5** and **6** in the **Appendix**.

Table 3: Hyperparameter Tuning Output on XGBoost

Parameter	Randomized Search Cross Validation	Grid Search Cross Validation
Learning Rate	0.1	0.1
Maximum Depth of Tree	200	200
Minimum Loss Reduction for Split	0.2	0.15
Percent of Columns for Sampling	0.7	0.8
Regularization L1 Penalty	10	5
Regularization L2 Penalty	10	5
Evaluation Metric for Validation Dataset	AUC	AUC

Using these parameters, the analysts ran the model and observed the ROC-AUC score as viewed in **Table 4**.

Table 4: ROC-AUC Score and Out-Of-Bag Error Score on XGBoost

Cross Validation Technique	ROC-AUC Score
Randomized Search CV	0.85
Grid Search CV	0.93

We can see from **Table 4** that the Grid Search cross-validation has a better ROC-AUC score. **Figure 1** shows the ROC-AUC curve for Grid Search cross-validation technique.

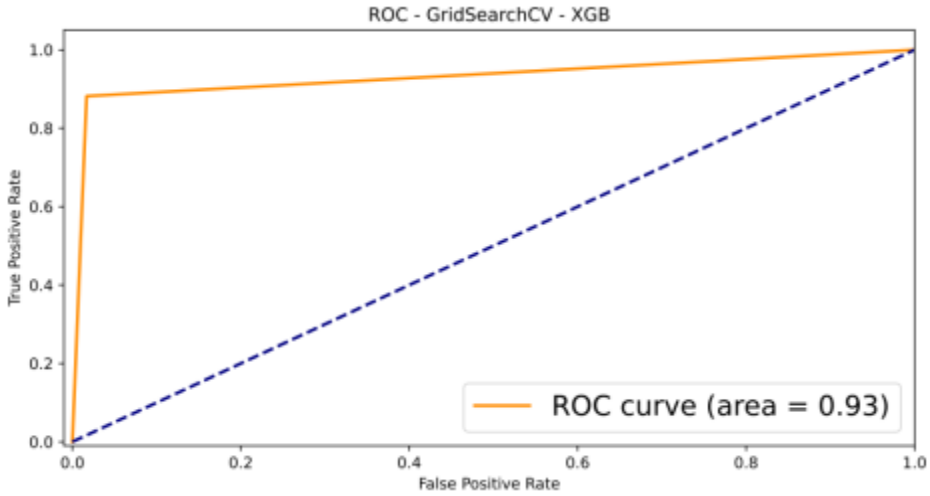

Figure 1: ROC Curve from the XGBoost Model

Results and Recommendations

While both models obtained similar results for the ROC-AUC metric, they approached the problem differently. One way to understand the differences between these models is by investigating the importance of the variables in each model. The XGBoost model considered the existence of a Checking Account to be the most important variable in predicting whether or not a customer will buy the variable annuity product. The RF model considered the customer's Savings Account Balance the most important variable, followed by the Checking Account Balance. Interestingly, the XGBoost determined the existence of a Money Market Account to be an important variable, whereas the RF model found the Money Market Account indicator variable to be less important than a random variable. Previous analysis for the Department determined the Money Market Account, or its corresponding account balance, to be an important variable. Furthermore, the RF model with Grid Search cross-validation found the random variable to be the fifth most important variable in the model, as shown in **Figure 5** in the **Appendix**.

The above observations and the ROC-AUC scores lead the analysts to conclude that the XGBoost is a better model to predict which customers will buy the variable annuity product. The analysts noted that the RF model technically achieved the highest ROC-AUC score of 0.94 with a Randomized Search cross-validation. However, in the RF model, the random variable which has no association with predicting whether a customer buys the variable annuity product, is ranked as more important than other variables known to be important in the business context. This is a signal of potential overfitting, and thus the team recommends proceeding with the XGBoost Grid Search model with hyperparameters shown in **Table 3** on page 3.

Conclusion

This report details the process of building a Random Forest model and an XGBoost model to predict whether customers of the Department will buy a variable annuity product. During data preprocessing, the analysts handled missing data, dropped highly correlated variables, and tuned hyperparameters with cross-validation techniques to find the best-fitted model. The ROC-AUC score of the final Random Forest Randomized Search model and the XGBoost Grid Search model is 0.88 and 0.93, respectively. For future predictions, the analysts recommend the Department proceed with the XGBoost Gird Search model since it identified important variables that meet business logic.

To better understand these models' ability to predict which customers will buy the insurance product, the analysts need to evaluate these models on unseen data. The analysts also recommend investigating why branches 14, 15, 18, and 19 are not offering certain products. Furthermore, it is important to keep in mind the tradeoff between the predictive power of the machine learning models and the explainability of the logistic regression model. Moving forward, the analysts suggest the Department consider trying more machine learning models as small improvements in predictive power may have large financial implications at scale.

APPENDIX

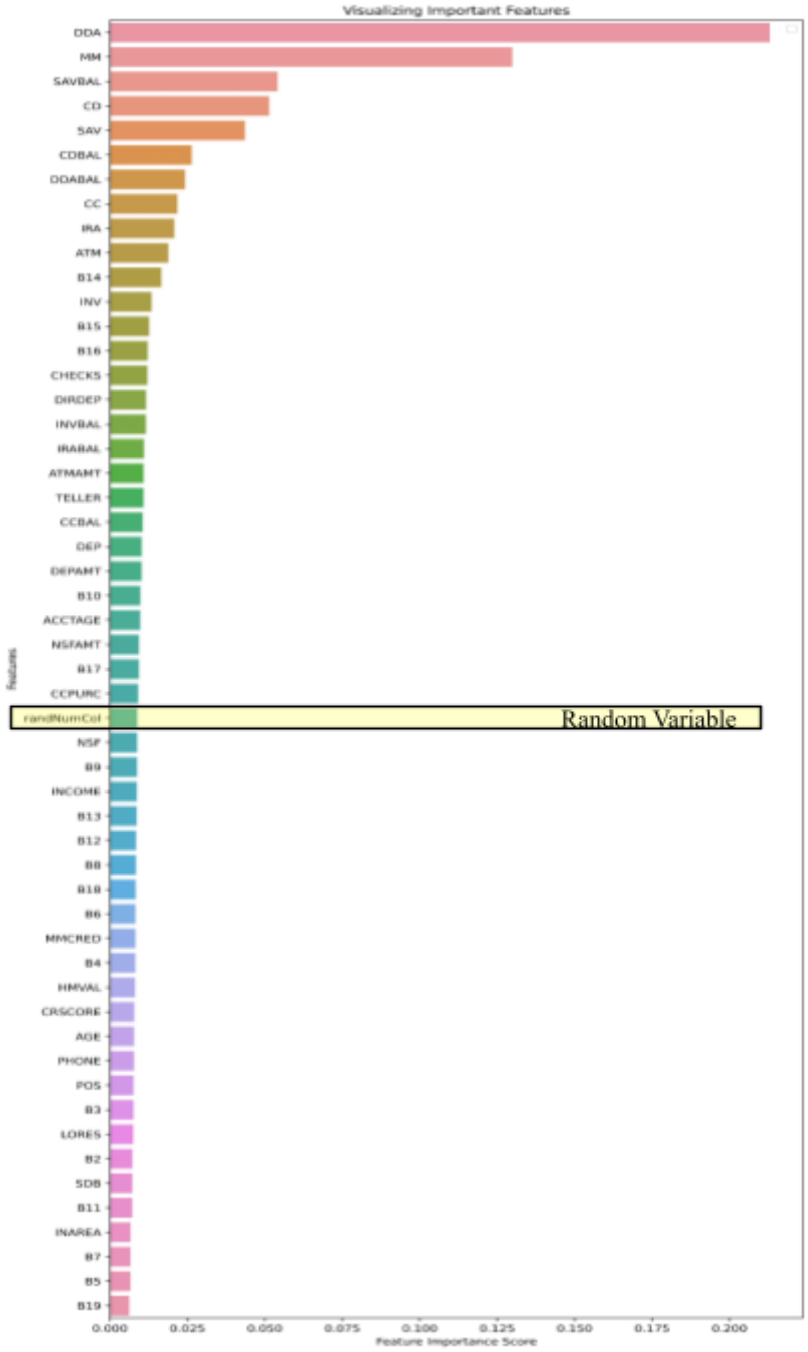

Figure 2: Variable Importance - Grid Search cross validation - XGBoost

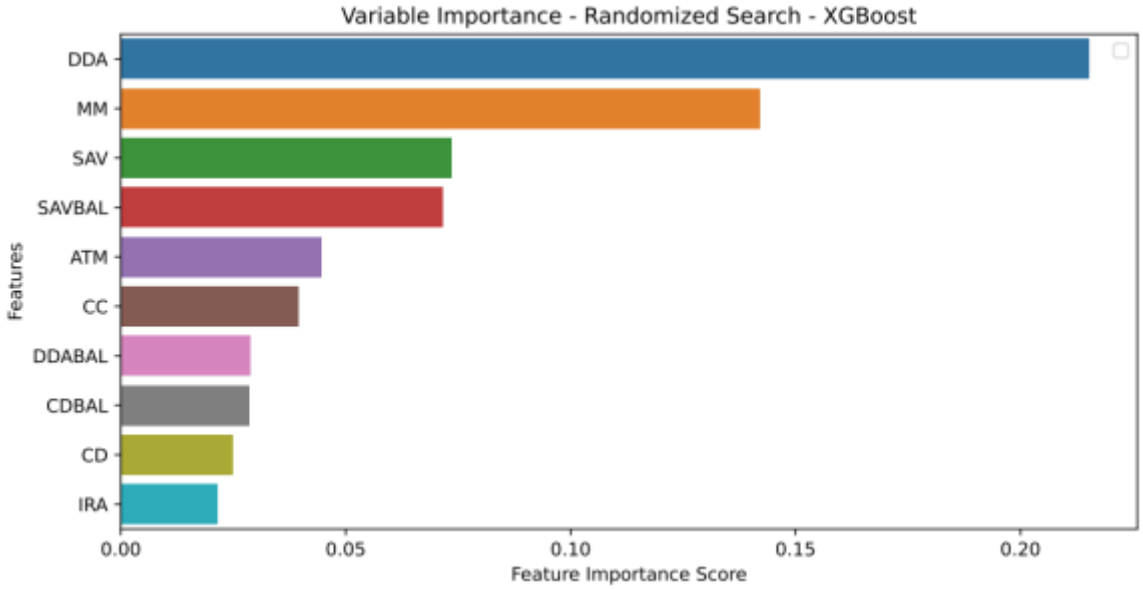

Figure 3: Top 10 Variable Importance - Randomized Search cross validation - XGBoost

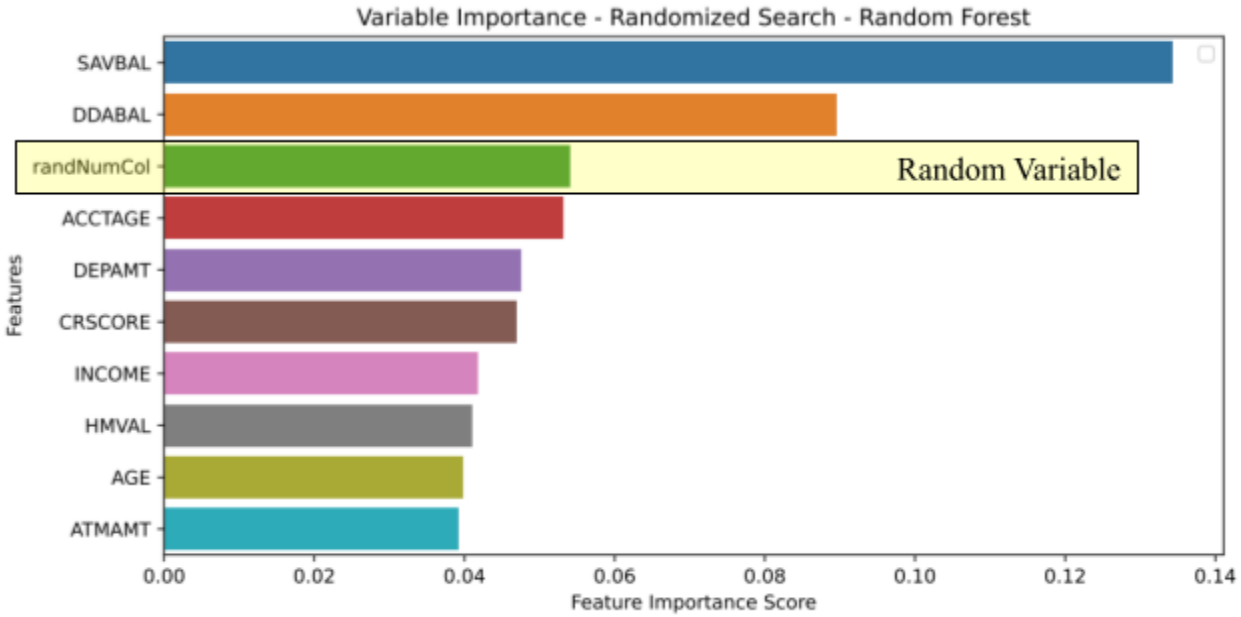

Figure 4: Top 10 Variable Importance - Randomized Search cross validation - Random Forest

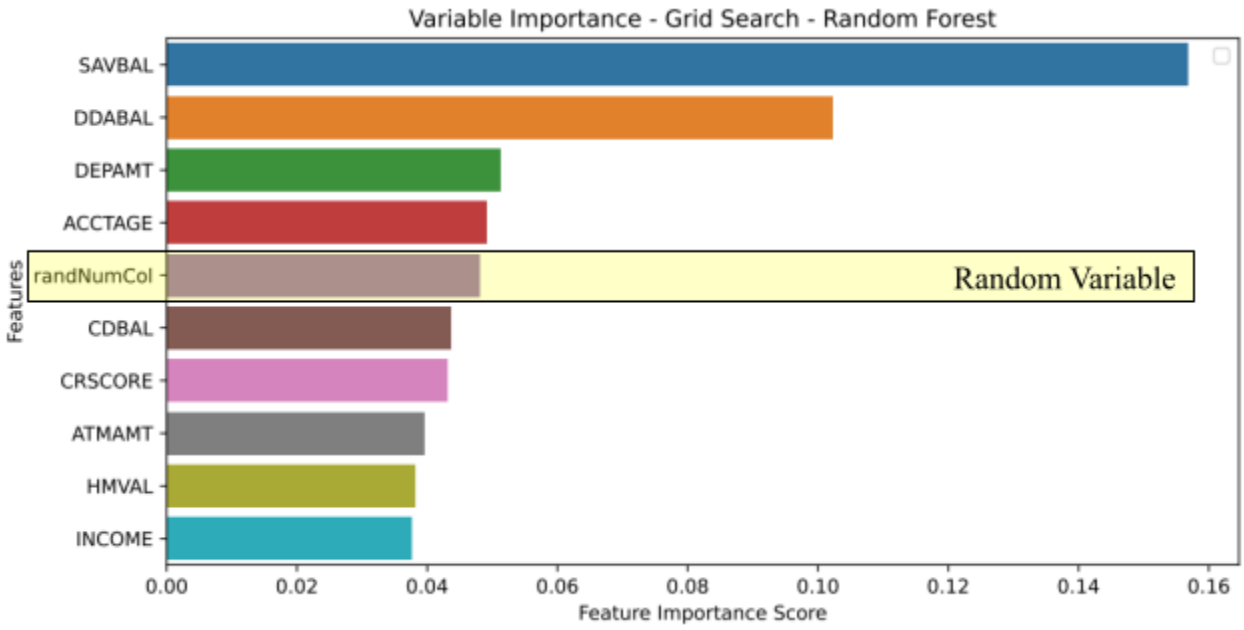

Figure 5: Top 10 Variable Importance - Grid Search cross validation - Random Forest

Homework Report Checklist

The team member(s) responsible for checking each item should enter their initials in the field next to each question. All items should be addressed before submitting the assignment with the initialed checklist attached.

Sections & Structure

Overview

BB	Is the overview concise?
BB	Does it provide context about the business problem? <Content>
BB	Does it briefly address your team's work, quantifiable results, and recommendations? <Action>
BB	Does it offer audience-centered reasons for recommendations? <Context>

Body Sections

BB	Does the report body include information on methods, analysis, quantifiable results, and recommendations?
BB	Is content grouped into appropriate sections (methodology, analysis, results, recommendations)?

Conclusion

BB	Does the report have a conclusion?
BB	Does the conclusion sum up the report and emphasize relevant takeaways?

Structure

BB	Does each major section have a heading?
BB	Are sections, subsections, and paragraphs organized logically for easy navigation?

Visuals

Introduction, Discussion, and Captions

NT	Is each visual introduced in the text before it appears?
NT	Is each visual close to where it is introduced?
NT	Does each visual include a title with the following information: type (table or figure), number, and a descriptive caption?
NT	Is each visual discussed and interpreted in the text?
NT	Are figures and tables numbered separately?
NT	Are table captions above the table? Are figure captions below the figure?

Visual Design

BB	Do figures/tables use audience-friendly labels rather than variable names?
BB	Are the visuals easy to interpret?
BB	Are the visuals appropriately sized?
BB	Do tables appear on one page (not split between 2 pages)?
BB	Are legends and axis labels included for figures?
BB	Are numbers in tables right aligned?

BB	Are the visuals designed well (ex: re-created in Word or Excel, not blurry or stretched,…)?

Document Design

Title Page Design

YT	Does it include a descriptive title?
YT	Does it state the team name, team members' names, and the submission date?

Table of Contents Design

EM	Does it list all the major sections of the report with corresponding page numbers?
EM	Do the page numbers and sections in the Table of Contents match the report?

Document Design for Entire Report

EM	Is a standard typeface (Calibri, Arial, etc.) used?
EM	Is the size of the body text between 10-12 pt.?
EM	Are headings and subheadings used to organize information?
EM	Are distinctive text styles (bold, italic, etc.) used to distinguish between heading levels?
EM	Are text styles for headings used consistently (ex: all level-one headings are bold)?
YT	Are all paragraphs an appropriate length (fewer than 12 lines)?
EM	Is white space used to indicate paragraph breaks?
BB	Are bullet lists used for a series of items and numbered lists to show a hierarchy?

Writing Style and Mechanics

Spelling and Capitalization

BKK	Are spelling errors located and corrected?
BKK	Is spelling consistent throughout (no switching between acceptable spellings)?
BKK	Is capitalization used appropriately (proper nouns, etc.)?
BKK	Is capitalization of words consistent throughout the report?

Grammar and Punctuation

EM	Are verb tenses used appropriately?
EM	Are marks of punctuation used appropriately?
EM	Is subject-verb agreement used in every sentence?
EM	Is the grammar checker updated and are underlined grammar issues addressed?

Writing Style

BKK	Are all sentences in the report easy for your audience to understand quickly?
BKK	Are most sentences written in active voice?
BKK	Are idioms and vague words eliminated from the report?
BKK	Are acronyms introduced before being used?
BKK	Are well-written topic sentences included at the beginning of each paragraph?
BKK	Are lists parallel?
BKK	Is the appropriate point of view used when addressing your audience or describing team actions?