CMC Banking insurance product

Orange 4
Erin Blake
Susan Davis
Keegan Donlen
Relwan Onikoyi
Jason Wang

October 21st, 2024

Table of Contents

CMC Banking Insurance Product

Overview

The Department of Customer Services and New Products at the Commercial Banking Corporation (hereafter the "Bank") seeks to identify which customers are most likely to purchase a variable rate annuity product. In the previous phase, we developed a logistic regression model to address this business challenge. We now extend our analysis by testing decision tree models.

We assessed recursive partitioning trees (rpart) and conditional trees (ctree) using performance metrics and a lift curve. We found the conditional inference tree model performs comparably to the logistic regression model in classification metrics but has a superior lift of 1.92 for the top 30% of ranked customers. Key variables that were particularly important in the conditional inference tree model were saving account balance (SAVBAL), checking account balance (DDABAL), indicator for checking account (DDA), indicator for money market account (MM), and indicator for certificate of deposit account (CD). By leveraging the key variables identified by the tree and the insights gained from the splits, the bank can more effectively allocate marketing efforts and budget to target customers most likely to increase their ROI.

Methodology & Analysis

This section of the report details our methodology for developing our decision tree and our analysis to determine the best tree.

Data Used

The dataset consisted of customer attributes for 10,619 individuals who were offered a variable rate annuity product. It included 48 variables: 47 describing customer characteristics and one indicating whether the customer purchased the product (INS). From the original dataset, 8,495 individuals were included in the training set, while the remaining 2,145 formed the validation set to assess our tree and regression models.

Model Development and Evaluation

Our team created two decision tree models, ctree and rpart, and evaluated their fit on the training set. To evaluate the effectiveness of our models in distinguishing between annuity purchasers and non-purchasers, we calculated both the coefficient of discrimination and concordance for the two decision tree models. Table 1 presents the performance metrics for both the rpart and ctree models.

Table 1: Performance Metrics for Decision Tree Models

Model	Concordance	Coefficient of Discrimination
Rpart Tree	69.53%	.15
Ctree	78.94%	.24

Table 1 shows that the ctree model achieved a higher concordance of 78.94%, indicating superior predictive power in identifying positive cases. It also achieved a higher coefficient of discrimination (0.24), suggesting it was better at differentiating between purchasers and non-purchasers of the annuity

product. However, the coefficient of discrimination graphs (see Figures 6 and 7 in the Appendix) reveal significant overlap between the two groups, indicating that both models struggle to distinguish them effectively. Both tree models show stronger performance in isolating individuals who did not purchase the annuity product from the overall pool of customers.

We also compared the ROC curves for both models to evaluate their performance (see Figures 4 and 5 in the Appendix). The ROC curve for the ctree rises sharply toward the top-left, indicating better predictive power. Visually, this suggests the ctree is more effective at distinguishing between purchasers and non-purchasers of the annuity product.

Results

This section of the report details the results of our chosen decision tree. After selecting the conditional inference tree for our model, we compared across metrics, as seen in Table 2.

Table 2: Conditional Inference Tree vs Logistic Regression Metrics

Metric	Ctree	Logistic Regression
Sensitivity	0.775	0.782
Specificity	0.649	0.657
Accuracy	69.3%	70.3%

Logistic Regression performed slightly better than the conditional tree in these metrics. We decided to use gains tables to better understand how these models rank the probability that someone will buy the insurance product. The results of the gains table for the conditional tree can be seen in Figure 1.

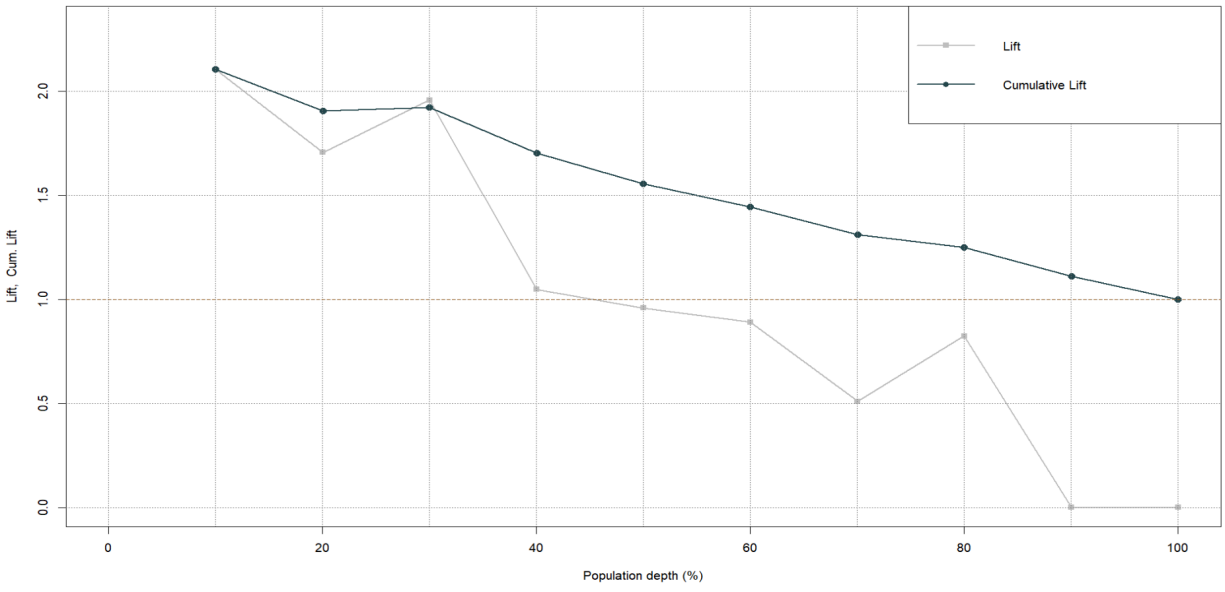

Figure 1: Gains Table for the Conditional Tree

Based upon the findings of Figure 1, the top 30% of the individuals selected by the model have nearly twice the odds of being an insurance product buyer than random selection. The logistic regression gains table can be seen in Figure 2.

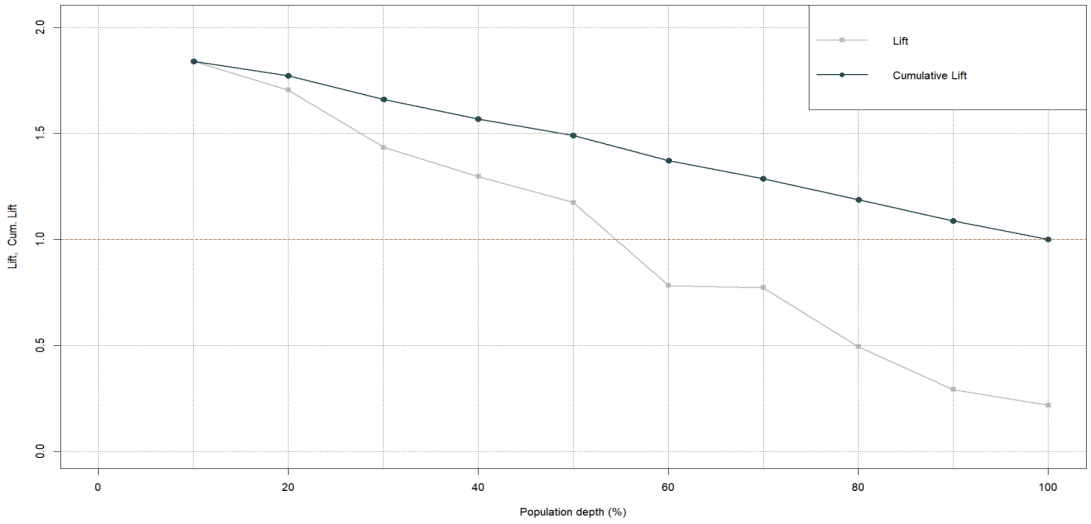

Figure 2: Gains Table for Logistic Regression

The gains table for the logistic regression model shows that the top 30% of those ranked by the model have 1.66 times greater odds than random selection. The conditional tree model shows clear superiority in identifying those likely to buy the insurance product for at least the top 30% of those ranked by probability.

The tree had 24 terminal nodes, the top three nodes were CD, SAVBAL, and DDA. For large decision trees like the ctree, visualizing it can be messy; we decided to visualize the importance of the variables used by the tree instead, as seen in Figure 3.

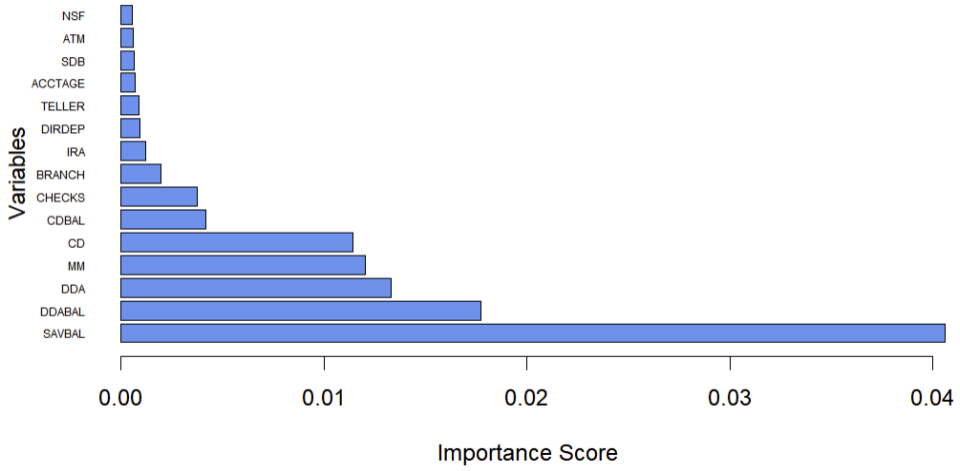

Figure 3. Importance Score Ranking of Conditional Inference Tree Variables

As seen in Figure 3, the five variables–CD, SAVBAL, MM, DDA, DDABAL– appear to be more critical in determining the final node for a customer. Most of these variable nodes appear near the top of the tree. DDABAL appears much lower, playing a vital role in many terminal nodes further down the tree. The ctree allows us to create certain customer profiles alongside their likelihood of purchase. For example, customers who have a Certificate of Deposit (CD) but have not set up direct deposit and have used an ATM exhibit a predicted probability of 85.9% of purchasing the variable rate annuity, indicating that this profile represents a highly likely segment for conversion.

Recommendations

After exploring and modeling with various decision tree approaches, we recommend the following:

1. **Utilize the insight from the tree splits to target customers for future marketing efforts**. Both the decision tree and logistic regression models tend to better isolate customers who are unlikely to purchase than those who are likely to do so. We recommend leveraging the insights from the tree splits and classification results to guide the allocation of marketing resources. As mentioned above, certain factors, such as savings and checking balances, play a role in influencing a customer's insurance purchasing decision. Customers with higher savings balances should be prioritized for greater marketing efforts and a larger budget than those with lower balances.
2. **Utilize the splits from the tree to bin the continuous.** For continuous variables like account balances, we suggest utilizing the tree's splits to create 'bins' that categorize these variables more efficiently and clearly while addressing the non-linear relationships observed in Phase 1. Binning can help uncover patterns that might be hidden when the data is left in its continuous format.
3. **Incorporate customer engagement variables.** We recommend incorporating data from marketing efforts aimed at customer engagement to enhance the understanding of the factors driving insurance purchases. Specifically, analyzing marketing channels, campaigns, content, and customer interactions can provide a more comprehensive view of the purchase decision. These upstream factors could influence a customer's buying choice but may not be captured by the current variables.

Conclusion

This report on the analysis of attributes related to variable rate annuity purchases shows the effectiveness of tree models alongside traditional logistic regression approaches. The ctree outperformed the recursive partitioning tree in performance metrics such as true positives and concordance. These are essential assessments that improve our ability to identify potential buyers.

While the logistic regression model shows a slightly better sensitivity and accuracy, the ctree's ability to segment high-potential customers offers a better strategic approach to targeted marketing. The gains table reveals that the top 30% of the identified customers have significantly higher purchase odds relative to random selection.

Our recommendations emphasize using these models to inform marketing strategies by targeting outreach based on classification outputs. Integrating previous marketing data will provide deeper insights into consumer behavior, allowing the bank to tailor its approach. By adopting these strategies, the bank can optimize marketing efforts, meet customer needs, drive growth, and enhance customer satisfaction by adopting the strategies listed in this report.

Appendix

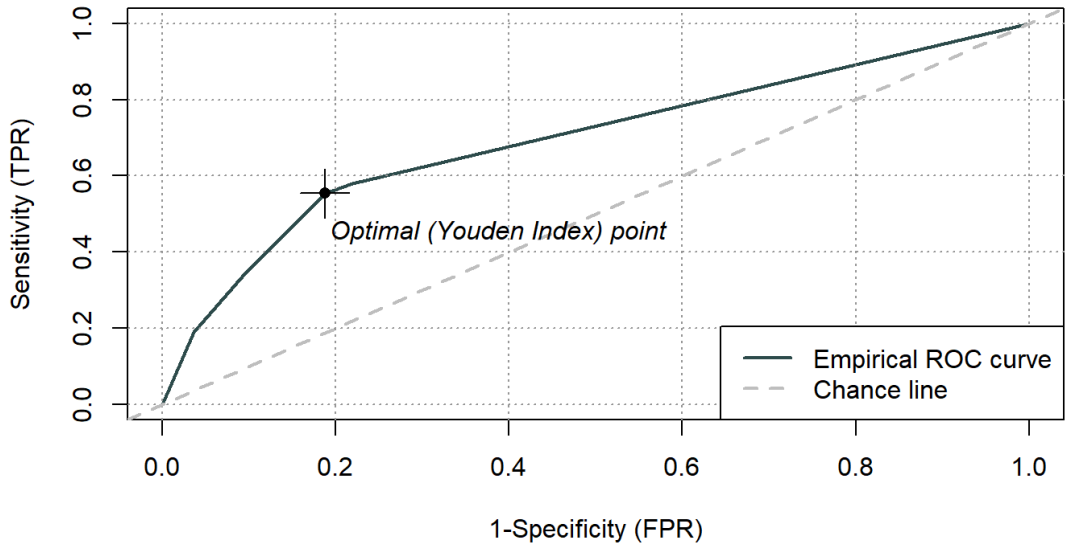

Figure 4: ROC Curve for Rpart Tree

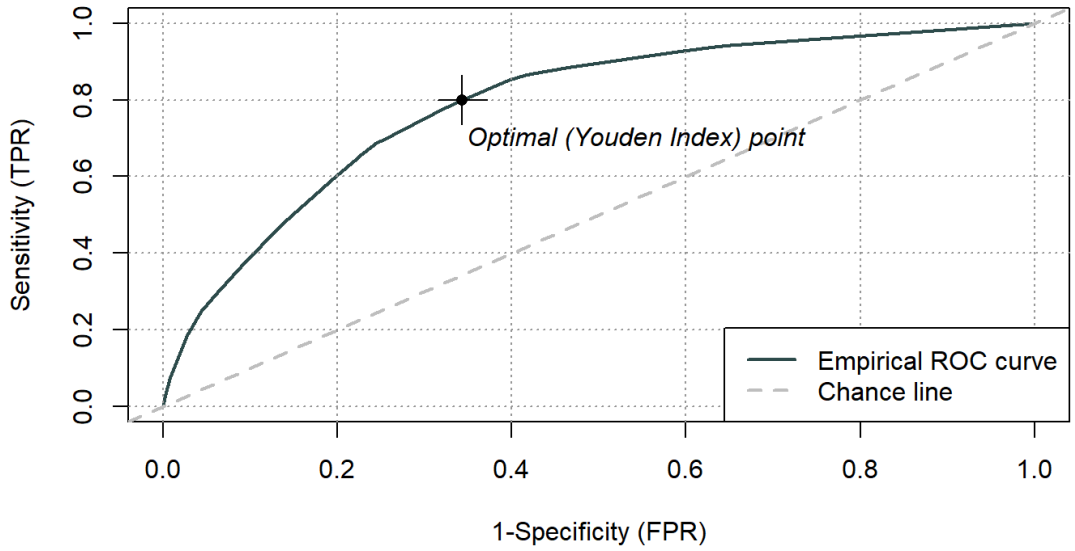

Figure 5: ROC Curve for Ctree

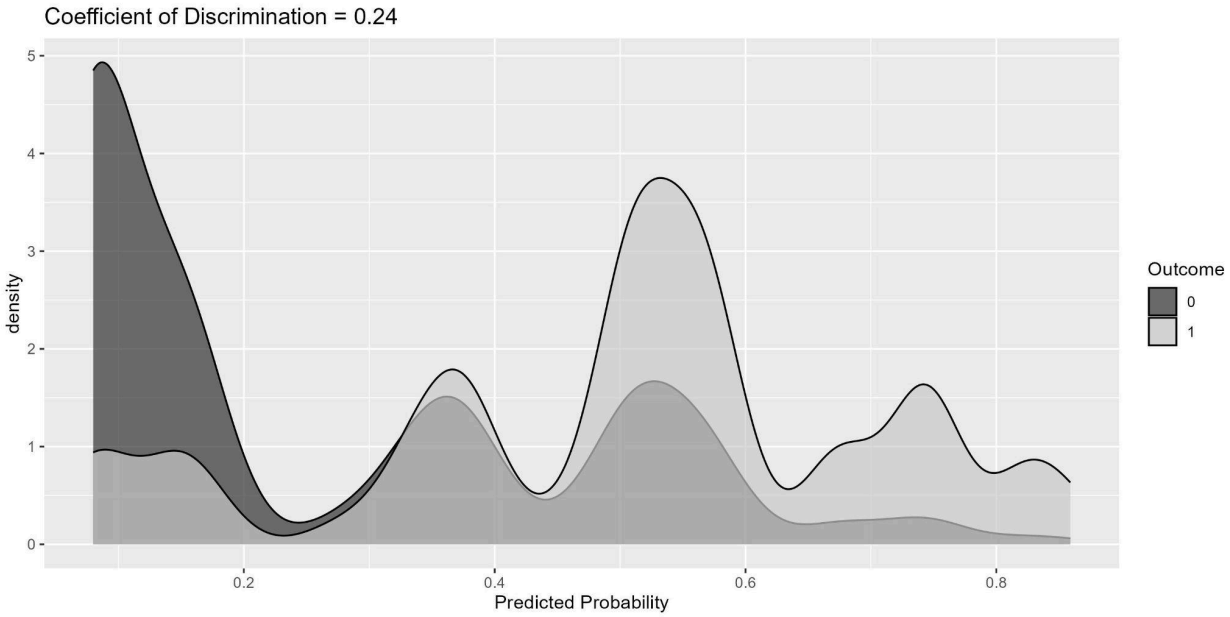

Figure 6: Coefficient of Discrimination Graphs- Ctree

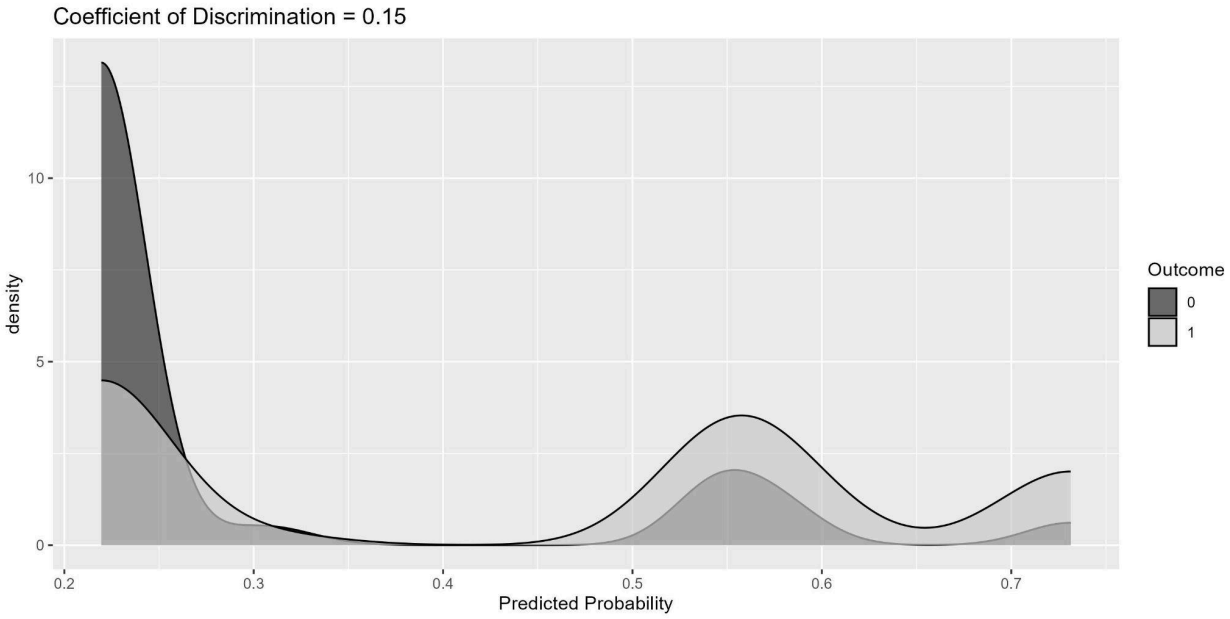

Figure 7: Coefficient of Discrimination Graph-Rpart tree

Homework Report Checklist

As instructed by Dr. Egan Warren, 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

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

Body Sections

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

Conclusion

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

Structure

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

Visuals

Introduction, Discussion, and Captions

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

Visual Design

RO	Do figures/tables use audience-friendly labels rather than variable names?
RO	Are the visuals easy to interpret?
RO	Are the visuals appropriately sized?
RO	Do tables appear on one page (*not split between 2 pages*)?
RO	Are legends and axis labels included for figures?
RO	Are numbers in tables right aligned?
RO	Are the visuals designed well (*ex*: *re-created in Word or Excel*, *not blurry or stretched*,…)?

Document Design

Title Page Design

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

Table of Contents Design

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

Document Design for Entire Report

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

Writing Style and Mechanics

Spelling and Capitalization

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

Grammar and Punctuation

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

Writing Style

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