BANKING INSURANCE PRODUCT

– Decision Tree Approach –

VOLT Analytics (Orange 12)
Tobi Yañez, Omry Brewster, Vishali Chauhan,
Tyler Farr, Lu Shen

October 19, 2023

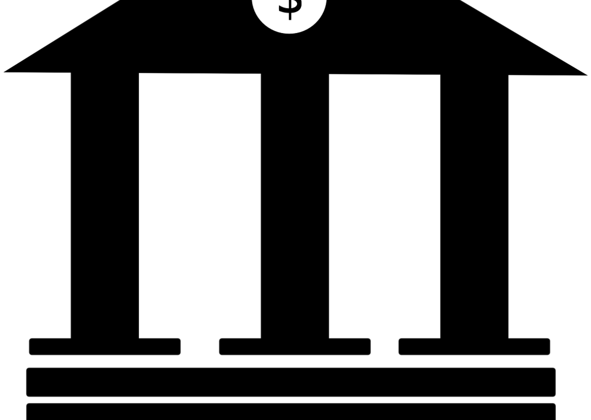


Table of Contents

Overview 1

Methodology & Analysis 1

Data Used 1

Decision Tree Development 1

Model Assessment 1

Results and Recommendations 2

Conclusion 4

Appendix 5

BANKING INSURANCE PRODUCT

Overview

The Commercial Banking Corporation ("the Bank") seeks to understand the customer profile of those likely to purchase their variable rate annuity product. VOLT Analytics has offered our services to create a decision tree model to assist the Bank in classifying customer habits.

According to our findings, our decision tree had a validation accuracy of 73.5%, which outperformed the previous logistic model with a validation accuracy of 70.2%. The Bank should utilize a decision tree model, as the higher accuracy provides the means to more effectively market to specific customer profiles and generate more revenue for the insurance product. We recommend focusing on customers with a savings balance of greater than $1,261 as a starting point.

Methodology & Analysis

Data Used

The data provided consisted of a training and validation dataset containing a shared set of 48 variables. The training dataset included 8,495 observations, while the validation dataset included 2,124 observations. From the training data provided, 35.7% of observations contained missing values. On the other hand, the validation had 35.5% of observations had missing values. However, this is not a problem, as decision trees can account for missingness. Although similar to the data provided for the previous project phase (RFP IP – F1.H3), the continuous variables here are left unbinned to allow the model to create unique cutoffs on its own.

Decision Tree Development

To create the best-fit decision tree model for our data, we assessed the Gini and information criteria when defining a splitting index. We found that the Gini criterion produced a lower misclassification rate on the training and validation datasets, so we chose this method to enhance our model moving forward.

Furthermore, we selected a maximum depth of 5 layers in our model to increase interpretability and limit overfitting. After performing initial trial and error with this argument, we found that a complexity parameter (cp) of 0.003 and a minimum of 130 observations as a cutoff for splitting yielded the most favorable results.

Model Assessment

To evaluate the model's efficacy, we calculated the Kolmogorov-Smirnov (K-S) statistic and the overall accuracy measure on the validation dataset. Moreover, we calculated variable importance to understand the individual effects of each variable originating from the interactions demonstrated in our model. These metrics were compared to our logistic regression model developed in RFP IP - H1.H3.

Results & Recommendations

Following a comprehensive analysis, we successfully created an optimized decision tree model, as illustrated in Figure 1, which assesses whether a customer is inclined to purchase the insurance product based on the listed cutoff points. Refer to Figure A.2 to review a complete list of variable definitions.

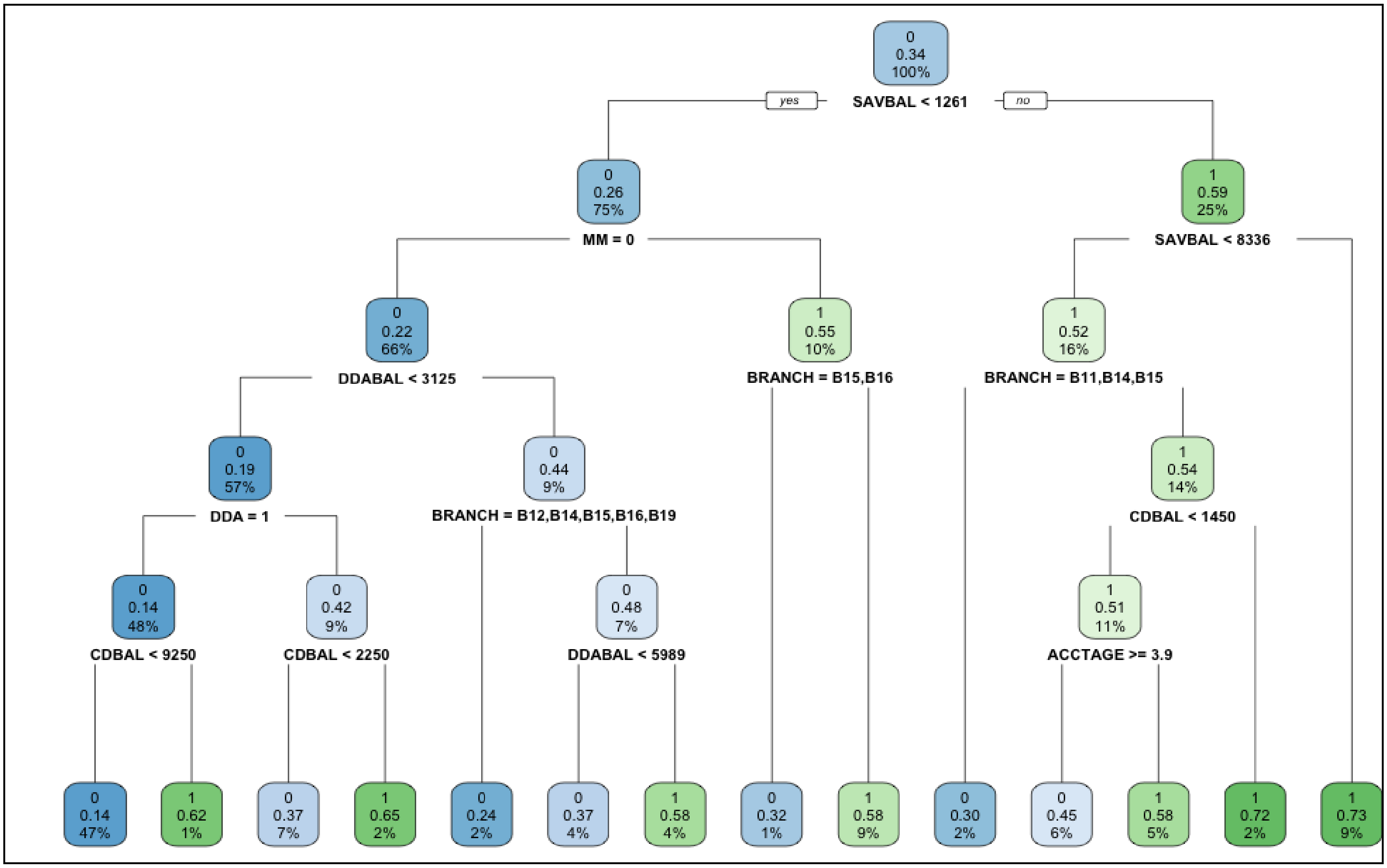


Figure 1: Final Decision Tree for Purchase of Insurance Product

The model visually distinguishes those customers who are likely to purchase the insurance product, represented by the color green, and those who are less likely to do so, denoted by the color blue. The construction of this model equips the Bank with the capability to analyze various customer attributes to forecast which customers are most predisposed to purchase a variable rate annuity product. We advise marketing to customers with a savings balance of over $1,261 as a starting point.

Within this decision tree, the variables that carry the most significance to the model tend to be positioned closer to the top, thereby underlining their importance in driving predictive accuracy. Figure 2 illustrates this fundamental concept of variable importance in a bar graph. The variables of heightened importance, such as “SAVBAL” (savings account balance) and “MM” (indicator for money market account) in the bar graph, align with those positioned at the top of the decision tree. This placement emphasizes the consistency in prioritizing these variables and reinforces their integral role in decision-making.

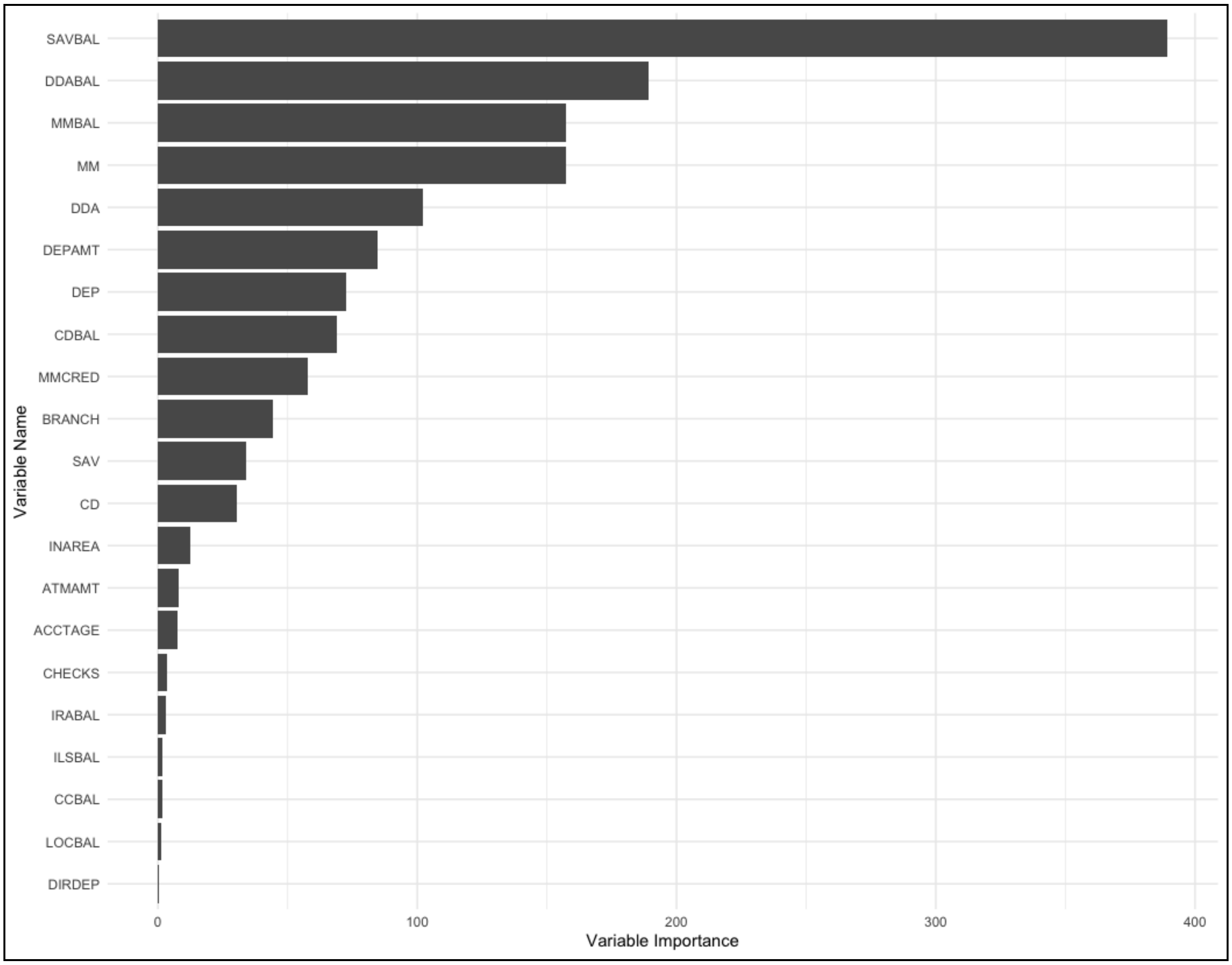


Figure 2: Variable Names Sorted by Importance

In the comparative analysis of the previous logistic and the decision tree models, we employed various statistical metrics to ascertain the most appropriate model for future predictions. Table 1 shows the accuracy and K-S statistic for each model.

Table 1: Statistical Values Between Models

Model	Accuracy	K-S statistic
Logistic Regression	70.2%	.47
Decision Tree	73.5%	.44

Upon review of the results presented, we have determined that the decision tree model should be the preferred choice for interpretation. Despite a slight reduction in the K-S Statistic value, this decision enables us to achieve a higher level of accuracy in our model while concurrently reducing misclassification by 3.3%. We recommend that the Bank utilize this to determine whether their customer will likely purchase a variable annuity product. Using our decision tree, the Bank can more accurately predict and efficiently market towards their customer segment compared to the logistic regression model.

Conclusion

Our team developed a decision tree model to predict customer inclination toward purchasing the variable rate annuity product. The model has a K-S Statistic value of 0.44 and an accuracy value of 73.5%. Compared to the previous logistic regression model, we decided to employ the decision tree model due to its higher level of accuracy and ease of interpretability.

Moreover, the decision tree model provides valuable insights regarding variable significance. Based on the decision tree model, savings account balance and checking account balance are the top two most important variables regarding the likelihood of the purchase. We recommend that the client allocate more marketing resources to those customer attributes identified at the top of the decision tree model, which can enhance the overall success of the marketing efforts.

Appendix

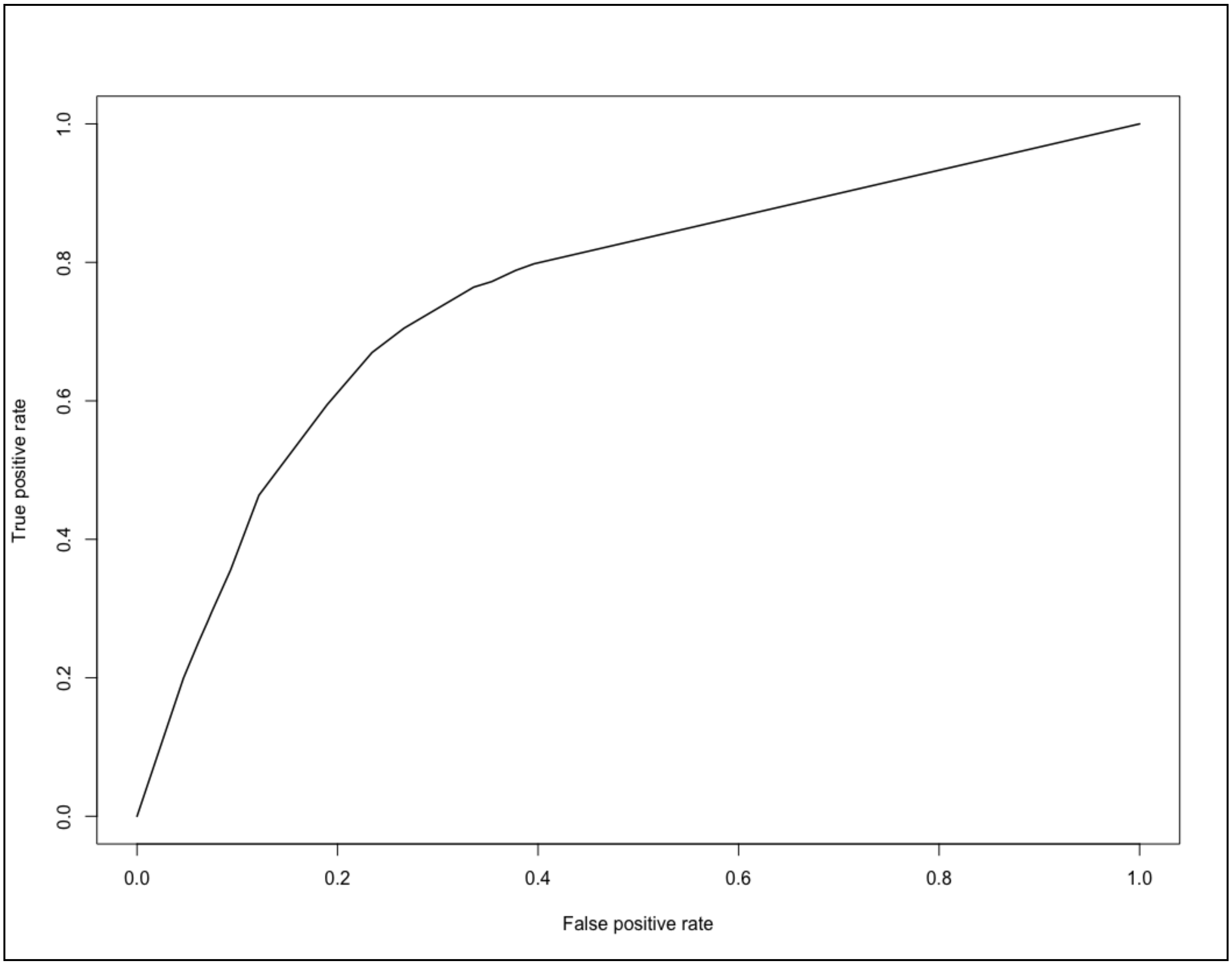
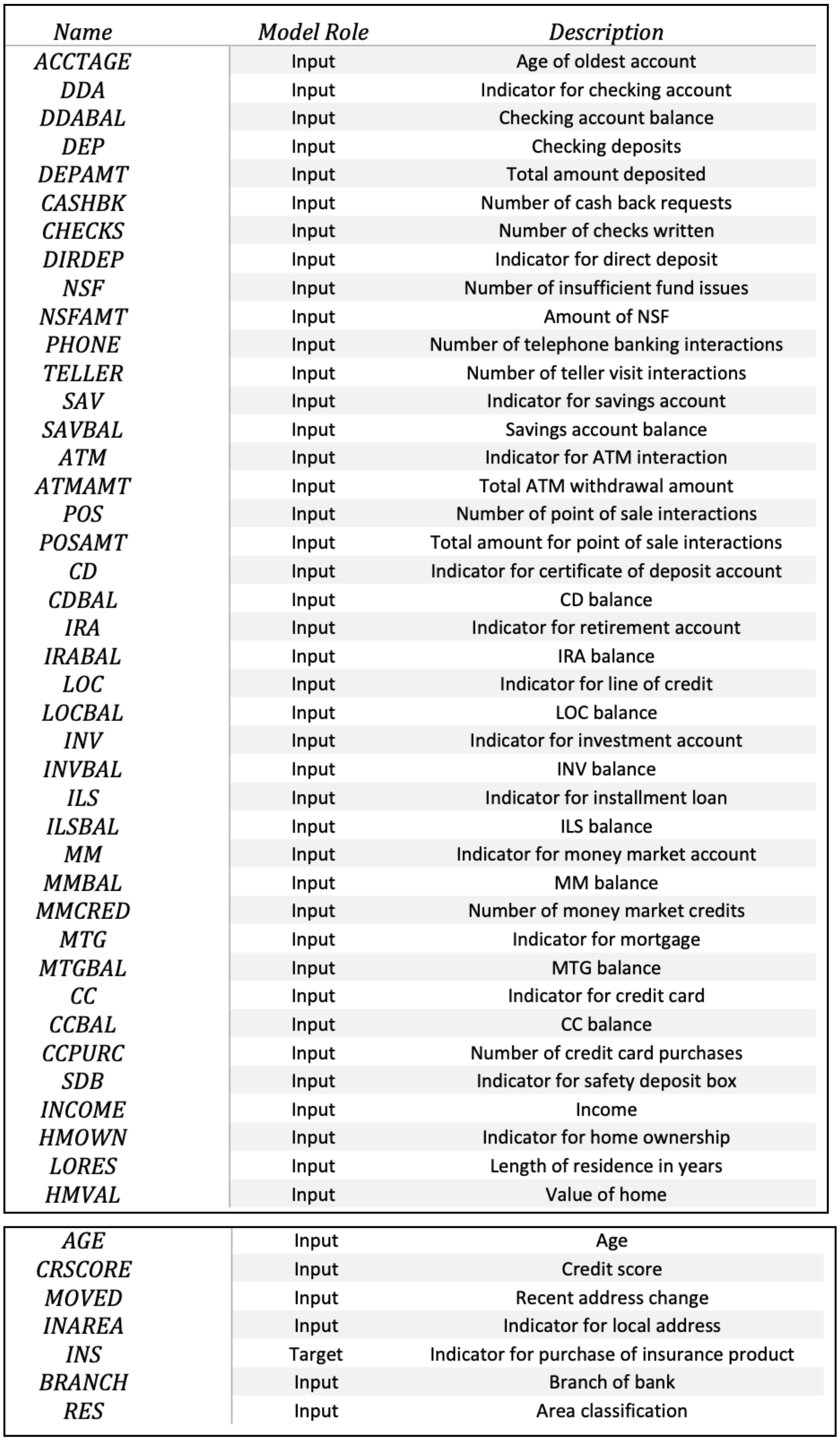


Figure A.1: ROC Curve of K-S statistic for Decision Tree Model



Name	Model Role	Description
ACCTAGE	Input	Age of oldest account
DDA	Input	Indicator for checking account
DDABAL	Input	Checking account balance
DEP	Input	Checking deposits
DEPAMT	Input	Total amount deposited
CASHBK	Input	Number of cash back requests
CHECKS	Input	Number of checks written
DIRDEP	Input	Indicator for direct deposit
NSF	Input	Number of insufficient fund issues
NSFAMT	Input	Amount of NSF
PHONE	Input	Number of telephone banking interactions
TELLER	Input	Number of teller visit interactions
SAV	Input	Indicator for savings account
SAVBAL	Input	Savings account balance
ATM	Input	Indicator for ATM interaction
ATMAMT	Input	Total ATM withdrawal amount
POS	Input	Number of point of sale interactions
POSAMT	Input	Total amount for point of sale interactions
CD	Input	Indicator for certificate of deposit account
CDBAL	Input	CD balance
IRA	Input	Indicator for retirement account
IRABAL	Input	IRA balance
LOC	Input	Indicator for line of credit
LOCBAL	Input	LOC balance
INV	Input	Indicator for investment account
INVBAL	Input	INV balance
ILS	Input	Indicator for installment loan
ILSBAL	Input	ILS balance
MM	Input	Indicator for money market account
MMBAL	Input	MM balance
MMCRED	Input	Number of money market credits
MTG	Input	Indicator for mortgage
MTGBAL	Input	MTG balance
CC	Input	Indicator for credit card
CCBAL	Input	CC balance
CCPURC	Input	Number of credit card purchases
SDB	Input	Indicator for safety deposit box
INCOME	Input	Income
HMOWN	Input	Indicator for home ownership
LORES	Input	Length of residence in years
HMVAL	Input	Value of home
AGE	Input	Age
CRSCORE	Input	Credit score
MOVED	Input	Recent address change
INAREA	Input	Indicator for local address
INS	Target	Indicator for purchase of insurance product
BRANCH	Input	Branch of bank
RES	Input	Area classification

Figure A.2: List of variable names and descriptions in dataset

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

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

Body Sections

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

Conclusion

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

Structure

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

Visuals

Introduction, Discussion, and Captions

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

Visual Design

VC	Do figures/tables use audience-friendly labels rather than variable names?
VC	Are the visuals easy to interpret?
VC	Are the visuals appropriately sized?
VC	Do tables appear on one page (*not split between 2 pages*)?

VC	Are legends and axis labels included for figures?
VC	Are numbers in tables right aligned?
VC	Are the visuals designed well (*ex*: *re-created in Word or Excel*, *not blurry or stretched*,...)?

Document Design

Title Page Design

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

Table of Contents Design

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

Document Design for Entire Report

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

Writing Style and Mechanics

Spelling and Capitalization

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

Grammar and Punctuation

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

Writing Style

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