**Optimizing Marketing Strategy**

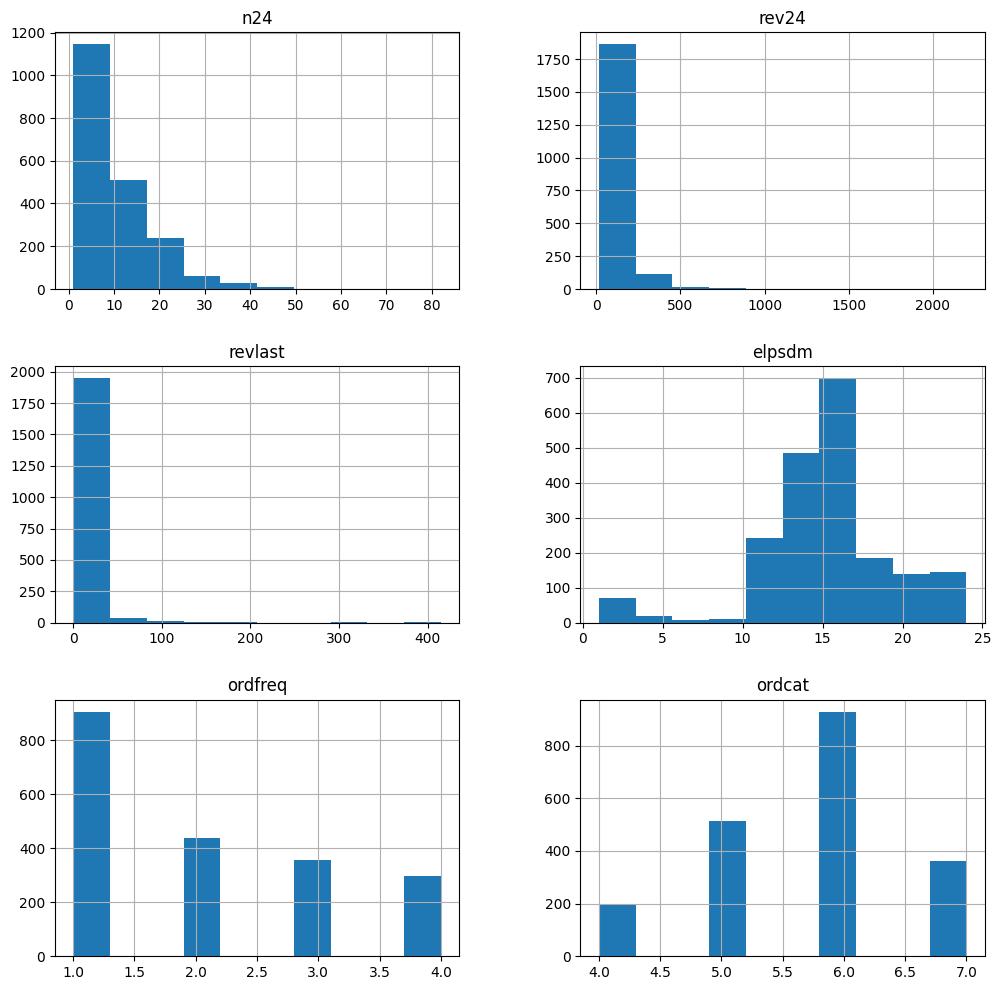
**BUAN 6383.001**

**Syam Menon**

Mihir Hirave (MDH230000)

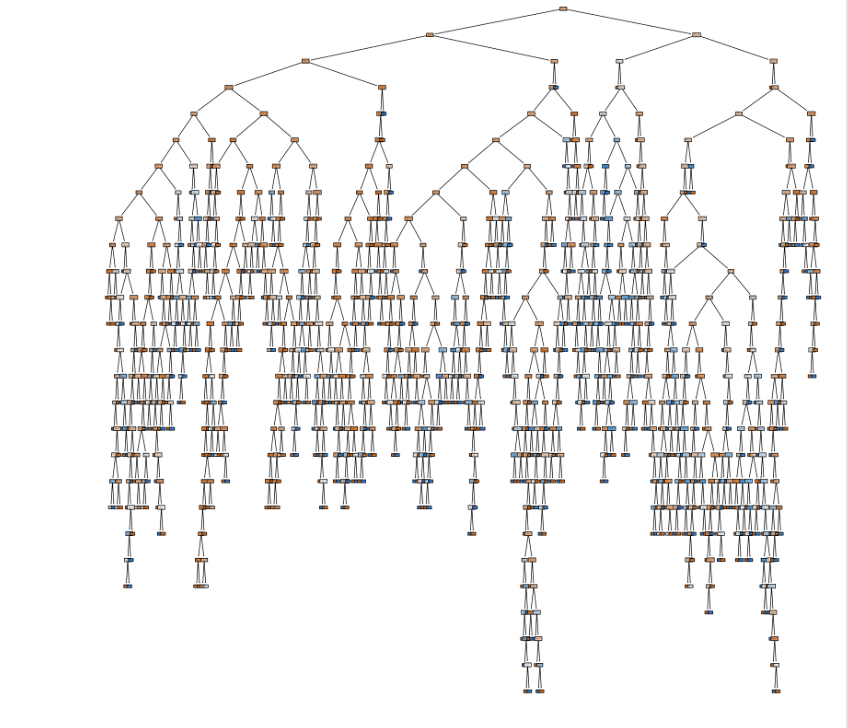
1. **Read in the data and review the non-binary variables to see if any are skewed and need to be transformed. If so, transform them and drop the non-transformed versions of the variables.** **Make sure that you do not include the customer identifier id in your calculations. Explain what you found, what transformations you applied, and why.**

We created histograms for all numeric variables. We found that n24, rev24, and revlast were strongly right skewed. Elpsdm was far less skewed, and ordfreq and ordcat were not skewed. We transformed the first three variables with a logistic transformation. This allows us to see the impact of percentage increase of each variable, rather than building models influenced by outliers.



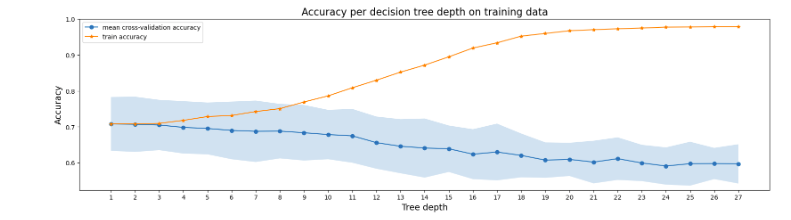
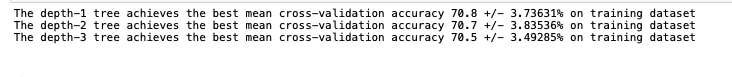
1. **Generate a decision tree on the entire dataset, without any limitations on the depth of the tree. Use entropy as the metric. What is the depth of the tree that is generated? Provide a plot of the tree.**

Our decision tree had a depth of 26. We found this depth to be quite deep and intensive to plot and understand.

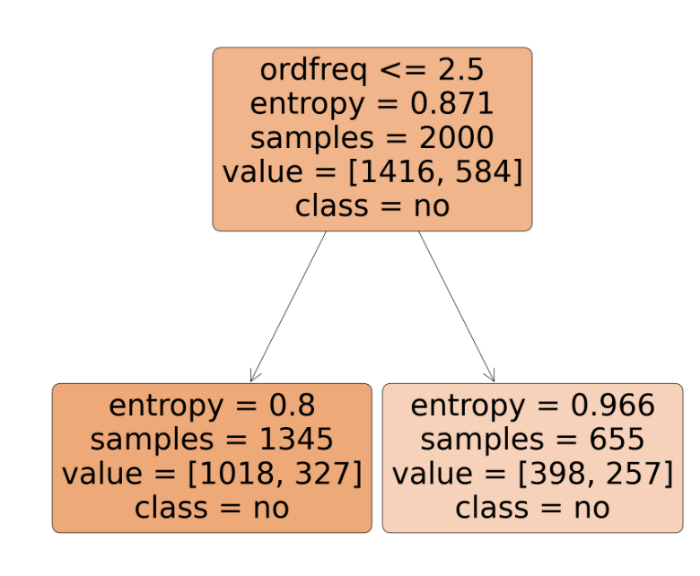


1. **We will focus on decision trees first, and try to identify the best decision tree classifier by pruning the tree at different depths. Use 10-fold cross validation and identify the best tree depth (again, using accuracy as the metric), by trying as many possible depths as you deem necessary. Provide your reasoning for using the values of tree-depth that you tried. Based on your results, what depth do you recommend? What is the accuracy associated with this tree depth? If you had to select the three best values of tree-depth, what would they be?**

We found that, while the total accuracy increased with additional tree layers, the mean cross-validation accuracy dropped as tree depth increased. We looked at the highest three values based on mean cross-validation accuracy in training. These values were a tree depth of 1, 2 and 3 with accuracies of 70.8%, 70.7%, and 70.5% respectively, which are our best and most recommended values for tree-depth. We chose depth 1 because it has the highest accuracy of 70.8%.



The figure below is the decision tree with depth 1. Here, it shows that ordfreq, is significant in influence the response class.



1. **Next, we will consider random forests. Develop a random forest classifier with 100 trees, using the three best values of tree-depth you identified in the previous question. Provide all relevant results. Which tree-depth results in the best random forest classifier? How does it perform relative to the best decision tree?**

Here, we used 10-fold cross-validation to be consistent and to compare with other models built in this project. We ran tests on the values of 1, 2, and 3 depths with 100 estimators in the random forest. We found that, in this version, depth 3 performed the best, with 71.2% accuracy.

This model performs slightly better with depth 3 as the accuracy is 71.2%. The accuracy of depth 1 is the same for decision tree and random forest. 

1. **Repeat this experiment with 50 trees. Provide all relevant results. Does your recommendation change?**

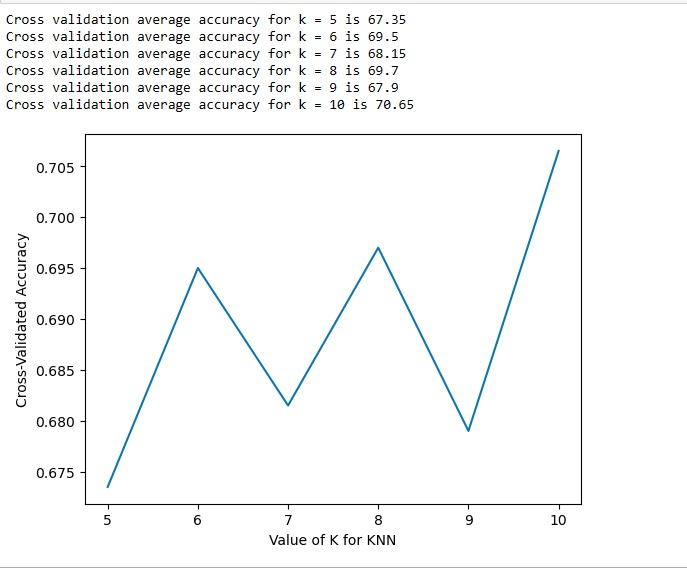
We performed the same test with 50 trees and received the same results. Depth 3 remains our recommendation with the highest accuracy of 71.20%. Since 50 and 100 trees give the same result, we chose the random forest classifier with 50 trees because it is simpler and would be more efficient to calculate.



1. **We will now consider *k*-nearest neighbor modes. Use 10-fold cross validation and identify the best value of *k*, by trying as many values of *k* as necessary. Keep in mind that very large values of *k* can affect speed, and that 5 is the default – try at least values from 5 to 10 (you can try more if you wish). Provide all relevant results. What value of *k* do you recommend? What is the accuracy associated with this value of *k*? If you had to select the three best values of *k*, what would they be?**

We have standardized the independent variables because KNN depends on distances between data points. Standardizing the data ensures that one variable does not have a greater impact than the others.

We performed a k-nearest-neighbor validation for values of k from 5-10. We recommend k values of 6, 8, 10 with accuracies of 69.5%, 69.7%, and 70.65% respectively. We chose k=10 since it has the highest accuracy among the others. Among the models built so far, this model performs worst in terms of accuracy.

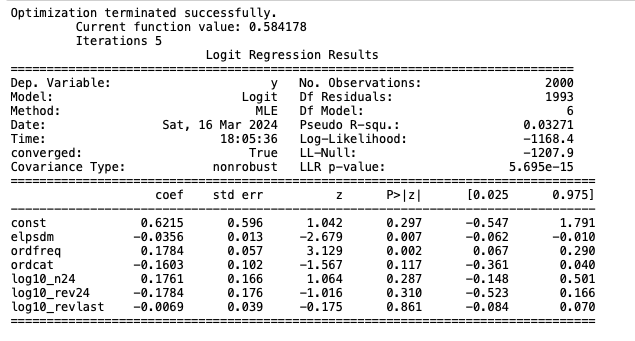


1. **Develop a logistic regression model using 10-fold cross validation. What is the associated accuracy?**

Our logistic model with 10-fold cross validation had an accuracy of 71.15%, which is has higher accuracy than other models expect for Random Forest with 50 trees.



1. **Develop a logistic regression model on the *entire* training dataset. Provide the output. What is the model developed?**



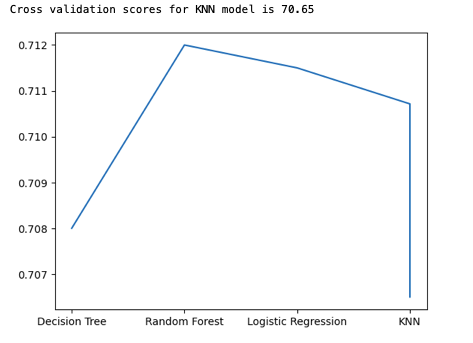
We developed a logistic regression model using the entire dataset. The log-likelihood function is –1168.4 and the LL-null value is –1207.9. Here, the LL-null is lower than the log-likelihood, which suggests that the logistic regression model provides a better fit to the data than the null model. The model suggests that elpsdm and ordfreq are two significant variables influencing response class.

1. **Using the four best models identified in each category (decision tree, random forest, *k*-nearest neighbor, logistic regression), perform an evaluation with 10-fold cross-validation. Your results should be similar (but not necessarily identical) to the results you have already obtained for these models. Across these four models, which one would you recommend, and why?**

We ran a 10-fold cross validation model on all models and compared their cross-validation scores.

We found that, in this version, all models performed between 70.8% and 71.2%. However, KNN (with 10 neighbors) performed the worst, and Random Forest model performed the best with 71.2%.





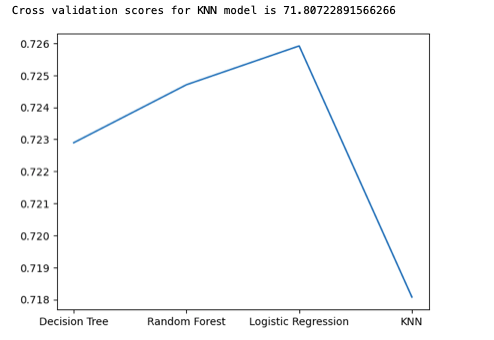
1. **Use the entire dataset to develop a final version of the recommended model for testing. Provide all details of the model (and the tree if the recommended model is the decision tree). What is that accuracy of this model on the training dataset?**

We fit the Random Forest model with 50 trees and depth 3 to the entire dataset and created a final recommended model. This model has an accuracy of 71.2%.

1. **If you were to focus on the “lapsing customers” (customers who made their last purchase 13 to 24 months ago), do you expect your model to be different? For the selected model, compare the quality of predictions for these customers relative to predictions for the others on records in the training set. Discuss your findings.**

We filtered the rows and created a dataset for the lapsing customer. We compared the accuracy scores of all the best models from each category (KNN, logistic regression, random forest, and decision tree) to find the model with best accuracy when it is trained using the lapsing customer data. We used 10 folds cross validation to evaluate our models, with accuracy as the scoring metric. We found that the best model for lapsing customer data is the Logistic Regression Model which is different from our earlier best model (Random Forest) as it gives the highest accuracy.





We created two disjoint datasets from our original dataset – One containing lapsing customers and the other containing non-lapsing customers. We already have the accuracy score for the lapsing customers from the previous part.

We evaluated our Logistic Regression model using the non-lapsing customer dataset using 10 folds cross validation. These are the results we got:

Lapsing Customers: Accuracy = 72.59%

Non-Lapsing Customers: Accuracy = 62.22%

Hence, our selected model provides us with better predictions when it is used on lapsing customers than with non-lapsing customers.

85% of the records in the training dataset are from lapsing customers, 15% of the records from the non-lapsing customers. So, the inferior performance of the logistic regression on the non-lapsing data does not significantly impact the performance of the model when it is applied to the entire training dataset.

From our findings, in Decision Tree classification, the highest accuracy is when the depth is 1. However, we cannot use this tree to make business decisions because it infers that no one is going to respond.



Using the best classifier model on the test dataset (Random Forest Classifier), we found that it predicted only 2 ‘1s’ out of 2000 i.e. 0.1% of the customers will respond to the mailers. Since most people will not respond to the mailers, the expected count of 0 as a response is significantly higher than 1. Thus, a model that predicts more zeroes is most likely to have high accuracy. This is one of the pitfalls of using accuracy as a metric to compare different models.