Data Analytics
INFSCI 2725
Homework 2
David Ball, Jaime Fawcett, Debdas Ghosh, Shruti Gupta, Piu Mallick

Introduction and Business Understanding

We have created a model that helps to predict the likelihood of future customers to churn within 6 months of expiration of their contract. We believe our model is a better model to implement because it accounts for overfitting – the model accounts for over-generalization to new data.

Data Understanding

The dataset we used includes churn data from MegaTelCo and contains 20,000 instances. We utilized the complete dataset for training, testing, and validating our model.

Data Preparation

To prepare our data, we needed to recode some of the variables to binary dummy values (e.g., 0,1). We did this for the target variable: leave (churn) = 1, stay = 0; and college: went to college = 1, did not attend college = 0. For the variables on customer satisfaction, usage, and likelihood to change plans, we split each variable into respective attributes associated with each possible choice. For example, satisfaction was split into five attributes: "very unsatisfied", "unsatisfied", "average satisfaction", "satisfied", and "very satisfied". Each of these attributes were assigned dummy variables 1 to represent the instance having the specific score, and 0 to represent the instance not having the specific score (see Figure 1). No instances of the dataset were removed. The final dataset featured 20,000 instances and 24 attributes, including the target attribute (leave).

long_calls	avg_calls	satis_average	satis_sat	...	usage_high	usage_low	usage_veryhigh	usage_verylow	plan_actlook
21.0	10.0	1.0	0.0	...	0.0	0.0	0.0	0.0	1.0
5.0	15.0	1.0	0.0	...	0.0	0.0	0.0	0.0	1.0
0.0	4.0	1.0	0.0	...	0.0	0.0	0.0	0.0	1.0
1.0	9.0	1.0	0.0	...	0.0	0.0	0.0	0.0	1.0
1.0	14.0	1.0	0.0	...	0.0	0.0	0.0	0.0	1.0

Figure 1. Sample of Cleaned and Recoded Data

Modeling

We settled on a classification tree model because our target variable (leave) is categorical. We developed a Python code script in Jupyter notebook to create our model. The final model had a depth of four and included seven of our 24 attributes: house, overage, income, leftover, handset, average calls, and long calls. These attributes were included given that they provided the highest information gain (Figure 2).

Evaluation

Overfitting means that the model is trained too closely to the training set but does not generalize well to data the model has not seen (test data). As a result, the model tends to be overly complex (contain too many variables) and tends to be highly accurate in predicting the value of the target variable for the training data, but inaccurate for new (test) data. To ensure our model was complex enough to obtain sufficient accuracy, but did not overfit, we created and evaluated various models with increasing complexity. We increased the complexity of our models by increasing the number of nodes in the classification tree model. To increase the number of nodes, we increased the depth of the tree, beginning with a depth of 1 and increasing to a depth of 15. This range allowed us to accurately assess when the model began to overfit and overgeneralize to the test dataset.

Data Analytics
INFSCI 2725
Homework 2
David Ball, Jaime Fawcett, Debdas Ghosh, Shruti Gupta, Piu Mallick

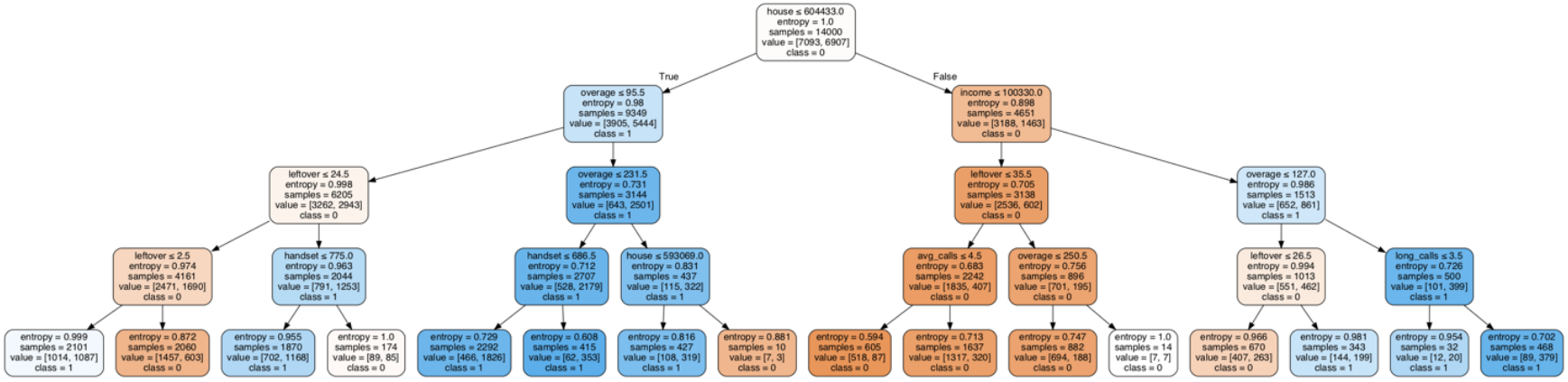

Figure 2. Final Classification Tree Model

For each model of increasing complexity, we evaluated the accuracy of the model using a holdout method. We divided our data into 70% training data and a 30% hold out dataset for testing the model. Each model was trained using the training data and then tested using the test dataset. We obtained accuracy scores for both training and test datasets for all 15 models. We used these scores to plot a **fitting curve** which allowed us to examine how the complexity of the model affected the accuracy of predicting churn for the training and test datasets.

Figure 3 shows how accuracy for the test data set begins to decrease once we reached a complexity with a tree depth of 5, even though the accuracy against the training dataset continues to increase. Therefore, the optimal complexity of the model was obtained with a tree depth of 4 (or 31 nodes), as in our final model. Our final model was **~70%** accurate for the training dataset and **~69%** accurate for the test dataset.

In addition to accounting for overfitting, we also wanted to ensure that our model's performance was sustained using different test datasets. It is possible that our model's accuracy was a result of a random occurrence based on how we divided our training and holdout datasets. To further evaluate our model's performance, we used a cross validation method utilizing 10 folds. The cross-validation method creates 10 "folds" of the dataset and tests the model 10 times, with each test predicting the accuracy for a unique sub-test dataset, thus providing 10 unique accuracy scores. Figure 4 shows the results of the cross-validation tests, compared to the original accuracy score for the test dataset of **69%**. From these scores, we computed an average accuracy score of **69.8%** and a variance of 0.00039 or **0.039%.** These statistics tell us that our final model does perform well against other datasets.

Data Analytics
INFSCI 2725
Homework 2
David Ball, Jaime Fawcett, Debdas Ghosh, Shruti Gupta, Piu Mallick

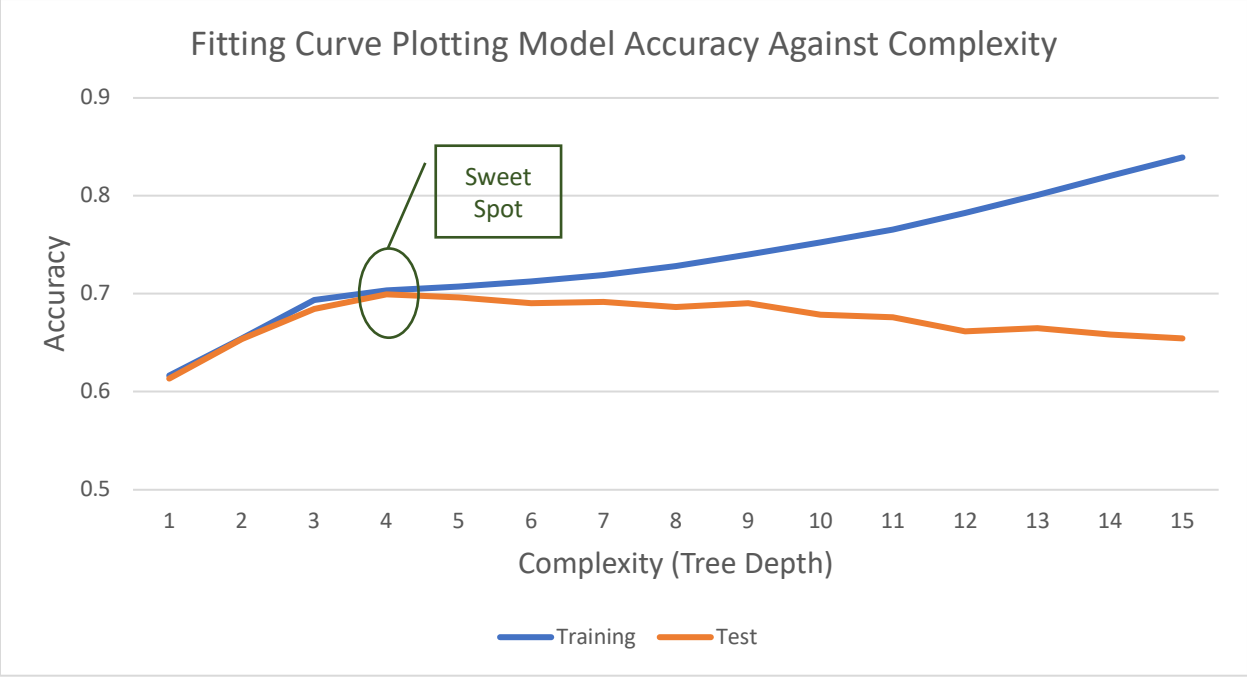

Figure 3. Fitting Curve Plotting Model Accuracy Against Complexity

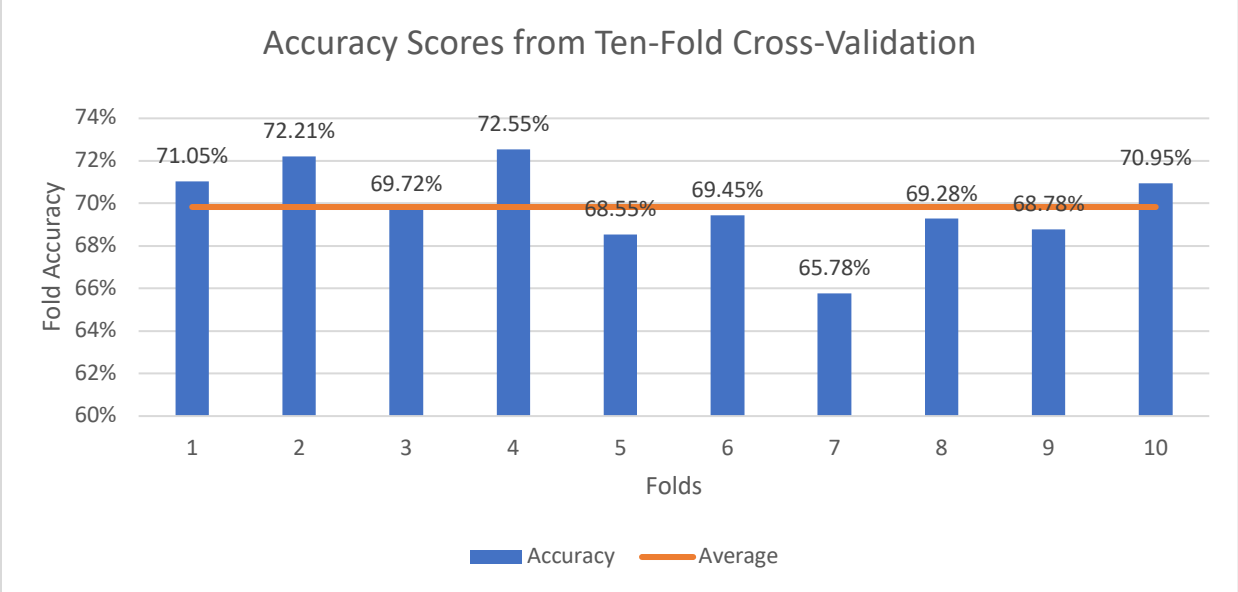

Figure 4. Accuracy Scores From Ten-Fold Cross Validation

Conclusion

In conclusion, our final model predicts whether a customer will churn within 6 months with approximately 70% accuracy. This has been validated through our fitting curve and cross validation methods which shows how alternative models, though they may be more complex and accurate for the training data, will provide less accurate predictions for new data.