Nick Iudiciani

MIST 4060 – Data Mining

Prof. Deokar

Assignment 1

**Part 1**

1. CRISP-DM, also known as Cross Industry Standard Process for Data Mining, is a 6-phase framework that provides guidelines for data miners. The phases go from understanding the business and its objectives, to data collection, to data preparation, then to the application of various data mining models, then to an evaluation stage which analyzes and applies the results, and finally to the deployment stage which organizes and presents the knowledge that is gained from running such models. SEMMA, on the other hand, is an acronym for Sample, Explore, Modify, Model, and Access. The stages go from sampling a large data set, to searching for anomalies and trends, to focusing data into a certain model, to the actual modeling of data, and finally into an evaluation for performance by tracking the model’s reliability. SEMMA is designed to provide solutions for identified problems within a business and is much like an incomplete version of CRISP-DM, because all 5 stages of SEMMA are very parallel to the 4 middle stages of CRISP, which are data understanding, data prep, modeling, and evaluation. CRISP-DM seems to offer more insights by allowing for business understanding and deployment, whereas SEMMA evokes more of an application that is used rather than to call upon more critical thinking.

Information found on <https://www.researchgate.net/publication/268770881_A_Comparative_Study_of_Data_Mining_Process_Models_KDD_CRISP-DM_and_SEMMA>

And <http://jesshampton.com/2011/02/16/semma-and-crisp-dm-data-mining-methodologies/>

1. A) Supervised learning: the data references a similar database, including the status of loans given out previously, so we can easily predict based on similar characteristics.

B) Unsupervised learning: though a trend may be identified, the recommendations are not tracked as being taken (or not taken).

C) Supervised Learning: since we know the status of other packets, it is easy to predict the status of our current packet.

D) Unsupervised Learning: Identifying segments of customers with similar characteristics has no known outcome, though after this is performed, supervised learning may be used for new customers to put into those segments.

E) Supervised Learning: we are predicting the financial status of a company using other company’s financial data.

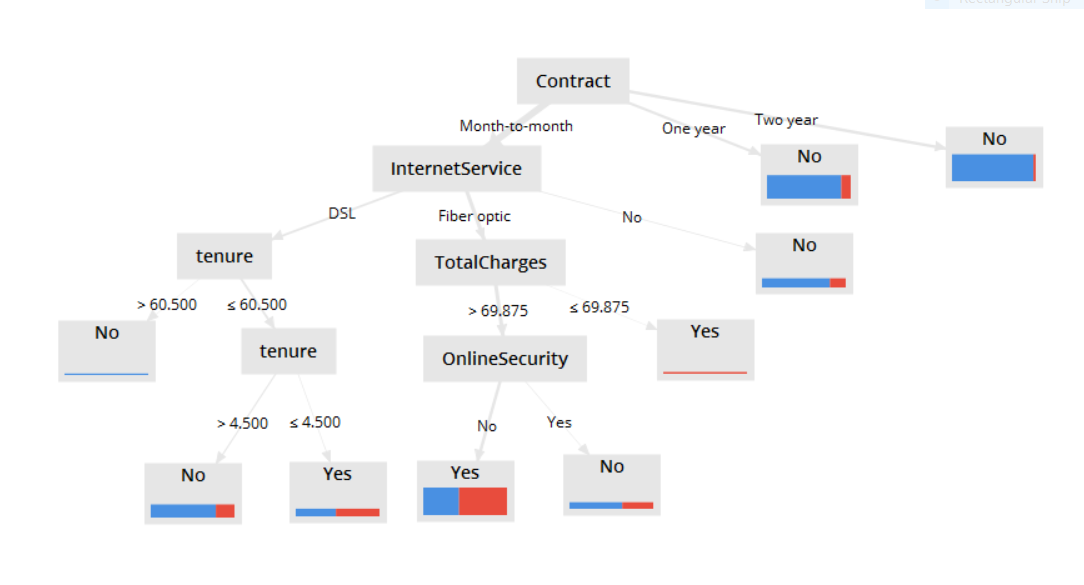
F) Unsupervised learning: we only have data relevant to the current situation. We cannot pull data from previous trouble tickets to predict how long it will take. Also, a trouble ticket may not describe all the problems that need to be examined.

G) Unsupervised learning: The mail system sorts mail automatically by the state’s zip code which are typically in order starting at 01001 (in Massachusetts, for example).

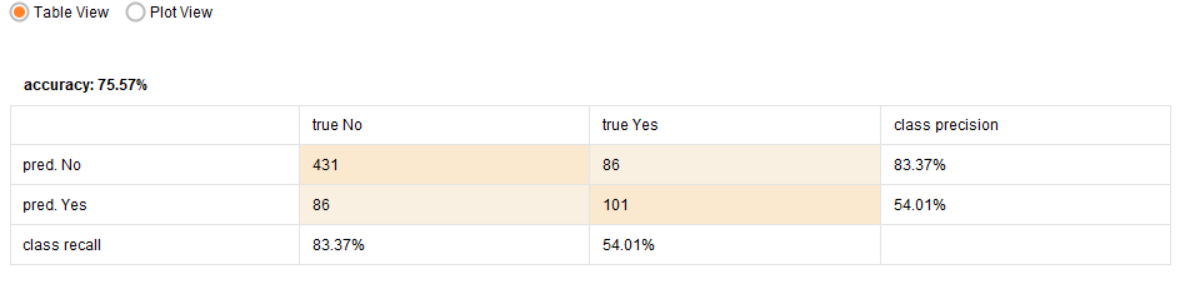
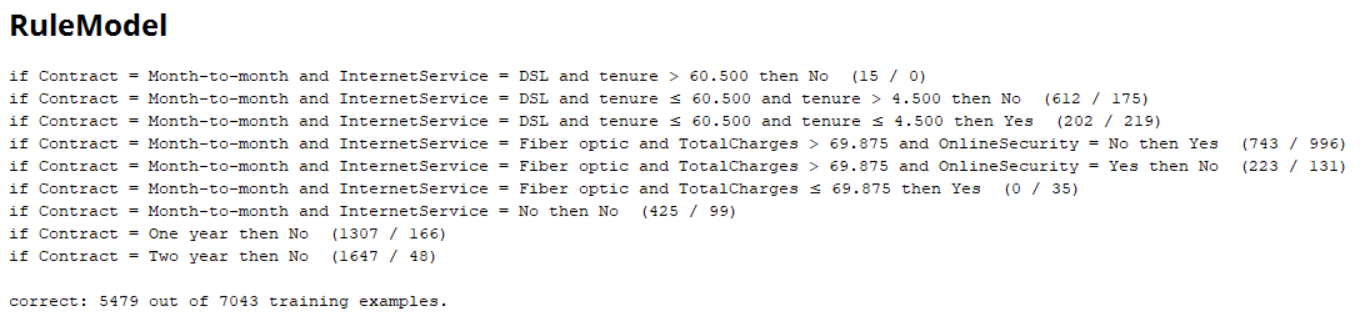
H) Supervised Learning: we are predicting that customers who buy certain items will come back for them again so we print coupons in order to further encourage their willingness to purchase from us again.

**Part 2**

*The resultant Decision Tree, Max Depth: 5, confidence 0.25, minimal gain 0.01, min leaf size 10 (TRIAL 1)*

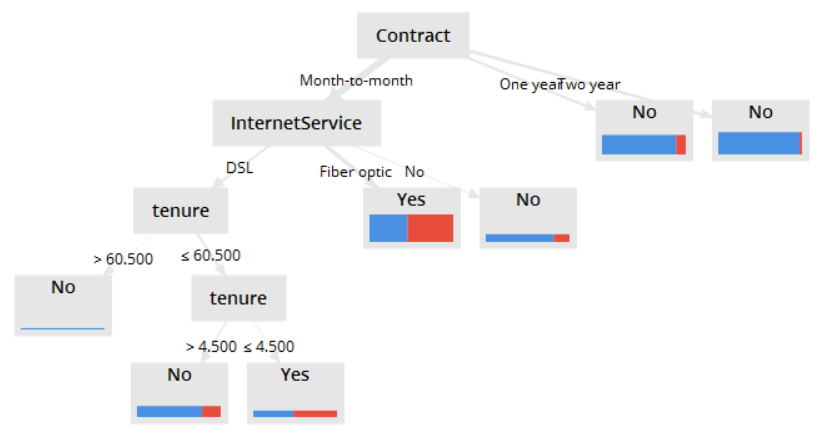


*Trial 1 Confusion Matrix and Rule Model*

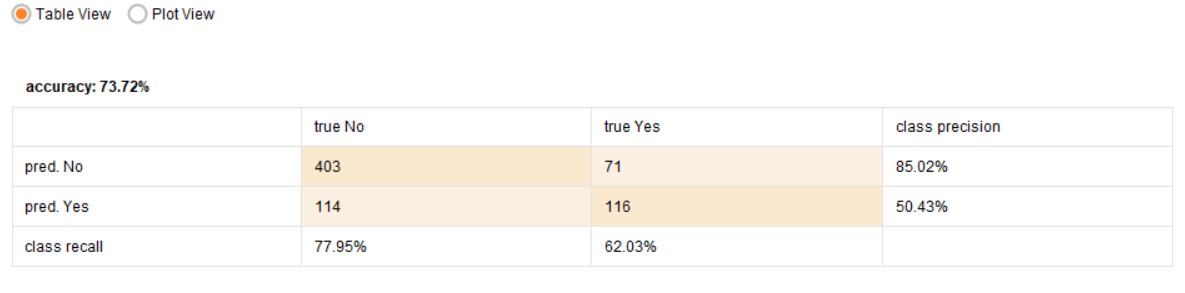


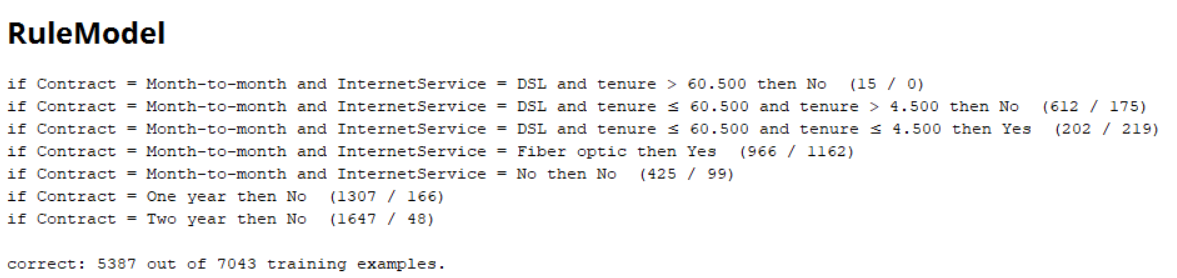
*Comments:* Looking at the resultant decision tree, a lower maximal depth (but not -1 since that is the value for a full tree with all attributes showing) of 5 allows us to see the portion of customers who churned based on their type of contract and internet service, their tenures, total charges, and if they had an online security program. Looking at the rule model, it seems that those who are locked into year-long contracts are more likely to stay with their service than churn, which makes up 2954 customers of the data set. It also turns out that more people decided to stay with the phone service if they were on a monthly contract for DSL, because the other options are paying for fiber optic which is currently very expensive, or not having internet service at all, which isn’t very popular in today’s world. The Confusion matrix however shows that the system was good at predicting who wouldn’t churn (83.37%) but not so much who would churn (54.01%).

*The resultant Decision Tree, Max depth: 6, confidence: 0.25, min gain: 0.04, min leaf size: 12 (TRIAL 2)*



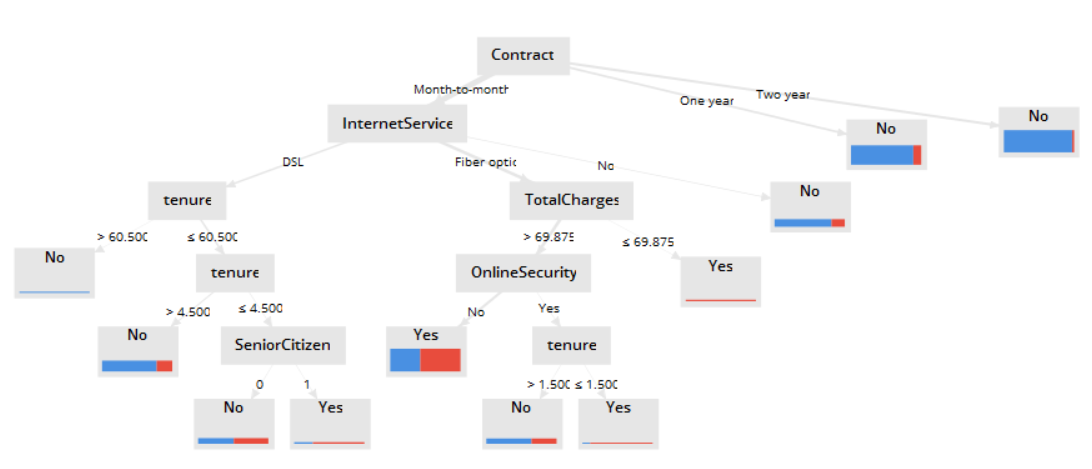
*Trial 2 Confusion Matrix and Rule Model*



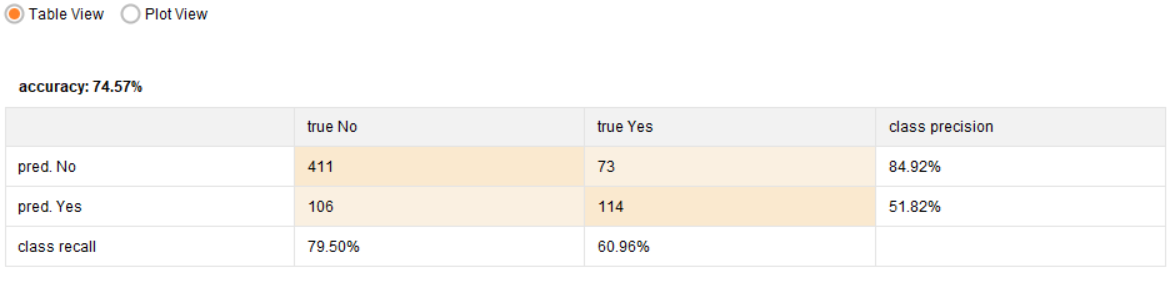


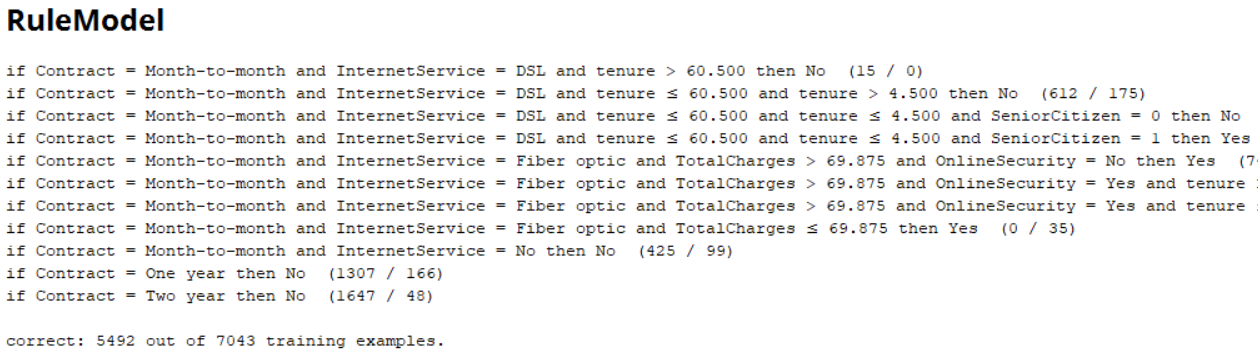
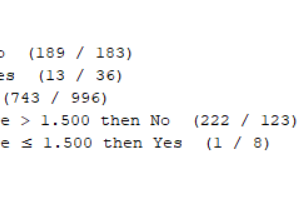
*Comments:* This decision tree has fewer leaves and node splits. This is due to the higher minimal gain of 0.04 as opposed to 0.01 in Trial 1. Thus, the rule model also reflects this change, which is the lack of a breakdown for customers with fiber optics or no internet service at all, merely showing their churn rate based on the type of service. The rule model shows that slightly more monthly subscribers would churn (1162) if they had fiber optics, perhaps because fiber optics is expensive. The confusion matrix shows different numbers than trial 1 as well. With slightly less accuracy (73% instead of 75% in trial 1), the system predicted almost 78% of the actual customers who wouldn’t churn *wouldn’t* churn, and predicting 22% would.

*The resultant Decision Tree, Depth: 10, confidence: 0.25, min gain: 0.02, min leaf size: 9 (TRIAL 3)*



*Trial 3 Confusion Matrix and Rule Model*





*Comments:* This decision tree would be much bigger if I changed the minimal gain to 0.01 instead of the current value of 0.02. This tree expands further on the tenure node, showing us the portion of customers who churned and were senior citizens. It also expands upon the churn rate for customers who had online security, relative to their tenures. One number, though a small one, occupied its entire pool of customers. The churn rate for customers who paid less than $70 a month for fiber optic was 100% in favor of churning (35 out of 35). Perhaps these customers felt that other customers were getting better quality service the more they paid. And these customers felt like they were getting excluded from a larger group of benefitting customers. The confusion matrix shows that out of the 517 that wouldn’t churn, 411 of them were true negatives (predicted no) and 106 of them were false positives (predicted yes).