

ISOM 3360 Assignment 2

Joshua Chang, Matthew Fung, XXX
{jchangad, yyy, xxx}@connect.ust.hk

1. Exploring and preprocessing the dataset

Before we build any models, we first need to explore and reformat the dataset. The dataframe consisted of 7032 Telco customers, and had 19 columns to each customer, with the last column (churn) being our target variable. This means the dataset had shape (7032, 19).

We checked and confirmed that the dataset did not contain any na values, and used `np.unique()` to find out the possible values to each column. After exploring the dataset, we are able to group the columns of the dataframe into three main categories:

- General information

These columns contained basic information that every customer had, for example their gender, whether they have a partner, their monthly charges and their contract length.

- Phone service related information

These columns had information on whether the customers were subscribed to their phone service, and whether they had multiple lines with Telco.

- Internet service related information

These columns had information on whether the customers were subscribed to their internet service, had online security, online backup and so on.

Since most columns contained only a few possible values, we converted the dataframe into a NumPy array for easier modelling later on. The detailed conversion rules can be found in the python notebook.

2. Building decision trees (Task 1, Question 1)

We built **FILL THIS IN LATER** decision tree models in total. Each model was trained and validated using 10-fold cross validation on 70% of the whole dataset. To ensure our trees are reproducible for fair comparison between different trees, we fixed “random_state” as “2211”, Evaluation of the trees were done using **accuracy, precision, recall**, and **AUC** as metrics, where each metric was taken as the simple average over all 10 folds. The summary of our models can be found in table 1.

Tree	Accuracy	Precision	Recall	AUC
Basic (§2.1)	0.735	0.0	0.0	0.0
Other (§xx)	0.0	0.0	0.0	0.0
Other (§xx)	0.0	0.0	0.0	0.0

Table 1. Comparison between trees

2.1. The basic tree

The first and most basic model we built was a basic decision tree with all the columns used for training. It resulted in a