

ISOM 3360 Assignment 2

Joshua Chang, Fung Ho Yin Matthew, Chan Shun Hang

{jchangad, hymfung, shchanax}@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 7,032 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. Decision Trees (Task 1, Question 1)

We considered 3 different methods of building decision trees 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)	0.738	0.505	0.517	0.668
Adj. params (§2.2)	0.794	0.624	0.566	0.831
Ensemble (§2.3)	0.790	0.551	0.500	0.680

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 features used for training, and all the default parameters `sklearn` provides. This means using “entropy” to decide which feature to split, splitting with the best feature, and allowing the tree to grow until all leaves were pure or contained less than 2 samples. During each split, the model considers *all* attributes.

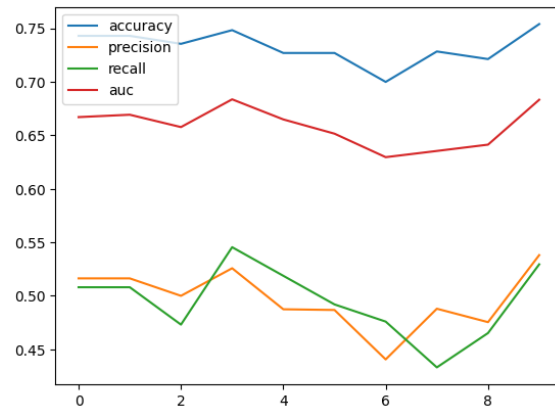


Figure 1. different metrics across the folds

Figure 1 shows the metrics across different folds during cross validation, and figure 2 shows the AUC curve. The metrics seem normal, however the AUC curve is too good to be true. Indeed, upon inspection we found that

```
model.predict_proba(X)
>>> array([[0., 1.], [0., 1.],
..., [1., 0.]])
```

which indicates that the model predicts the probabilities for

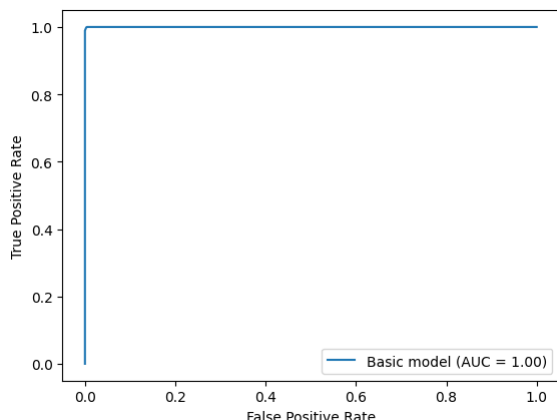


Figure 2. AUC curve for the basic tree

each sample with 100% certainty. In fact, when comparing it with the ground truth, out of 4,992 samples it correctly classifies 4,889 of them. This means adjusting the threshold won't affect the prediction results, which explains the absurd AUC curve.

The model also had a depth of 34 with 1,467 leaves. The reason the model overfits the dataset is because by default, `sklearn` will continue splitting until each leaf is pure or contains 2 or less samples, so these are the first parameters we are going to change in our next model to be built.

2.2. A tree with adjusted parameters

The next model we will build aims to solve the issue above. Specifically, we will adjust the following parameters, and find the best model by `GridSearchCV` using 10 folds.

- `max_depth`

Search range is `np.arange(5, 35, 5)`.

We chose this range because the overfitted model in section 2.1 had a depth of 34, so setting `max_depth` to 30 should be sufficient.

- `min_samples_split`

Search range is `np.arange(10, 110, 10)`.

We chose this range based on empirical testing. Even when we set lower / higher values, the best estimator found by `GridSearchCV` mostly took on values in this range.

- `min_samples_leaf`

Search range is `np.arange(10, 110, 10)`.

We chose this range based on the same empirical testing as `min_samples_split`.

Criteria	Accuracy	Precision	Recall	AUC
Accuracy	0.799	0.646	0.529	0.831
Precision	0.794	0.624	0.566	0.831
Recall	0.796	0.673	0.451	0.834
F1	0.794	0.624	0.566	0.831

Table 2. The best trees based on different criterion

After deciding how to build the tree, the next step is to decide which metric `GridSearchCV` should use to decide the best estimator. We used maximizing accuracy, precision, recall and F1 as our different criterion, and summarized the results in table 2 and in figure 3. As usual, each model was trained with `random_state = 2211`, and evaluated using 10 folds CV.

Our empirical testing shows that regardless of the criterion chosen, the model's performance is already a significant improvement compared to the basic model in section 2.1.

Out of the 4 criterion chosen, figure 3 shows that the models generally performed equally, so which out of the four models to choose will depend on the costs of misclassifying the customers, to be handled in the later sections. For now, we will pick the fourth model (F1 score as criteria) as the benchmark for this subsection, since it offers a balanced consideration of both precision and recall.

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

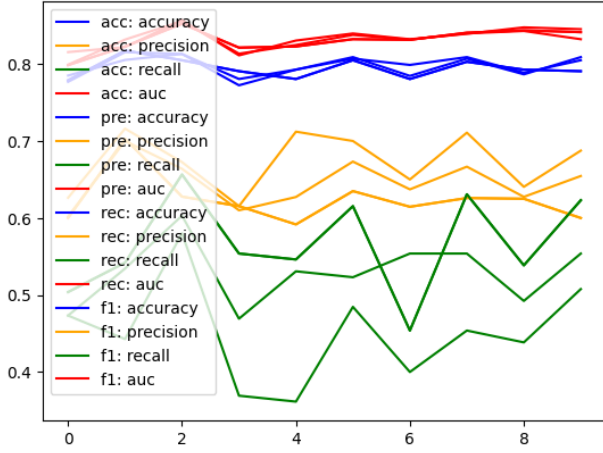
2.3. Ensemble trees

Recall in section 1, we discussed that the customers could be broadly grouped into two categories, those with phone service and those with internet service.

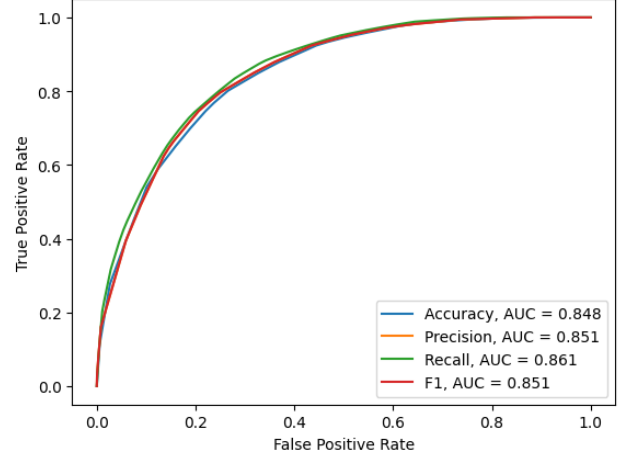
It follows from intuition that these two types of customers would have different reasons for continuing or discontinuing their contracts, so one idea to build the decision tree would be to first split the customers into three categories: (1) those subscribed to phone service only, (2) those subscribed to internet service only, and (3) those subscribed to both services.

Unfortunately, `scikit learn` does not have an easy way to specify the first split, so we have to do it manually. We built three models, corresponding to the three types of customers mentioned in the paragraph above. The results are summarized in table 3.

Note that since the sample size for internet only users is 456, when we did 10 fold CV we encountered a division by 0 error due to the small fold size. To make the comparison fair, we did not use CV to evaluate the models. Instead, we split off another 30% from the training set, and used that to validate our models.



(a) the performance of the four models across folds



(b) AUC curve for the four models

Figure 3. Performance of the four models in §2.2

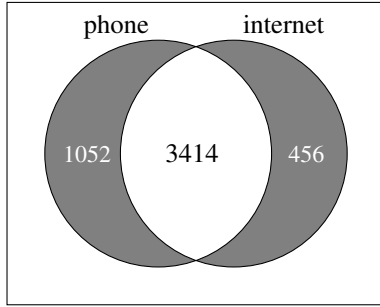


Figure 4. subscription service provides a natural split

Tree	Accuracy	Precision	Recall	AUC
Phone & Internet	0.753	0.623	0.610	0.716
Phone only	0.921	0.333	0.190	0.582
Internet only	0.766	0.520	0.394	0.639

Table 3. Comparison between ensemble trees

We then used the following formula to convert the score back to the whole training set, to mimic the effect of setting the first split¹:

$$\text{Score} = \frac{1052 \times P + 3414 \times P \cap I + 456 \times I}{4922} \quad (2)$$

¹This method does not make too much sense for the AUC metric, but we tried it on the whole training set and the result was similar.

Tree	Accuracy	Precision	Recall	AUC
Ensemble	0.790	0.551	0.500	0.680

Table 4. Ensemble tree performance

where

- *Score* denotes the specific metric in question (accuracy, precision, recall, AUC),
- *P* denotes the metric obtained from the model trained using only phone subscribers,
- *P ∩ I* denotes the metric obtained from the model trained using phone and internet subscribers, and
- *I* denotes the metric obtained from the model trained using only internet subscribers

The results of the Ensemble model are summarized in table 4.

2.4. Choosing the best model

Based on table 1, we choose the adjusted parameters model in section 2.2 to be our most effective model.

Recall that before we started training the 3 different types of models, we split off 30% of the dataset to be used as our testing set. Now that we have chosen our final model, we can use the testing set to evaluate the effectiveness of our model.

The results on the testing set are summarized in table 5. The results are similar, indicating that overfitting was unlikely.

Figure 5 shows the relative feature importance for the attributes that contributed to reducing entropy in the leaf

Tree	Accuracy	Precision	Recall	AUC
Adj. params (§2.2)	0.787	0.604	0.602	0.729

Table 5. model performance on testing set

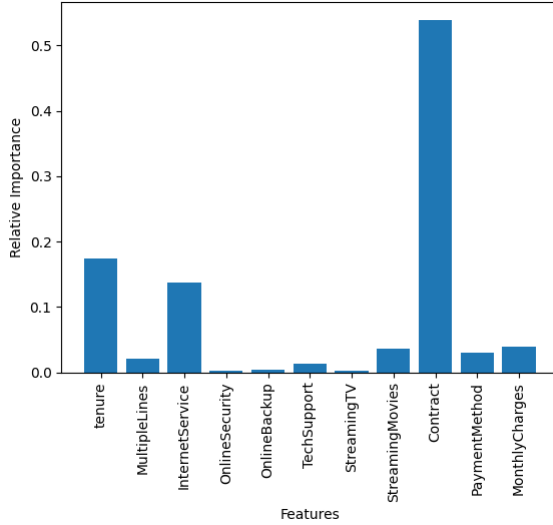


Figure 5. the relative feature importance for various attributes

Contract	Churn	Total	Churn Ratio
Monthly contract	1655	3875	42.7%
Yearly contract	166	1472	11.3%
Bi-yearly contract	48	1685	2.85%

Table 6. proportion of customers that churn, sorted by contract type

nodes. The factor that by far contributed the most was the attribute “contract”, which had three possible values: (1) Month-to-month, (2) One year, and (3) Two year .

We then tried to find out which contract type of customers were most likely to churn. The results are shown in table 6.

This result makes sense and matches our intuition - if customers signed monthly contracts, they can easily cancel their Telco subscription and switch to another service provider at the end of every month. On the other hand, the longer the contract, the less likely the customers would cancel their contract, as cancelling a longer contract requires a larger penalty for breaching the contract, and once they miss their contract cancellation window they will be automatically renewed for another yearly or bi-yearly period.

Regularization	Accuracy	Precision	Recall	AUC
None (§3.1)	0.803	0.654	0.544	0.841
L_1 (§3.2)	0.0	0.0	0.0	0.0

Table 7. logistic regression model performances

3. Logistic Regression (Task 1, Question 2)

We build two logistic regression models, one with no regularization and one using L_1 regularization. To ensure fair a comparison, both models were

- trained with the dataset normalized,
- trained with `random_state = 2211`,
- trained until they converged (tolerance $< 10^{-4}$) and
- fit with an intercept.

We report the accuracy, precision, recall and AUC score in table 7.

3.1. No regularization

The model’s parameters are shown in table 8.

The loss across folds is shown in figure 6, and the AUC curve is shown in figure 7.

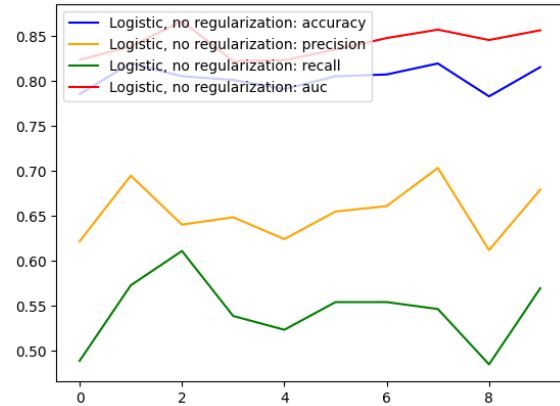


Figure 6. model performance across different folds

In general, the logistic regression model performed marginally better than the best decision tree model in section 2. Besides having a slightly lower recall, the accuracy, precision and AUC score were slightly higher. However, as we know from section 2, only 11 out of the 18 attributes contributed to reducing the impurity in the leaf nodes. So, we expect that using regularization would help us improve our model performance.

3.2. Lasso regularization

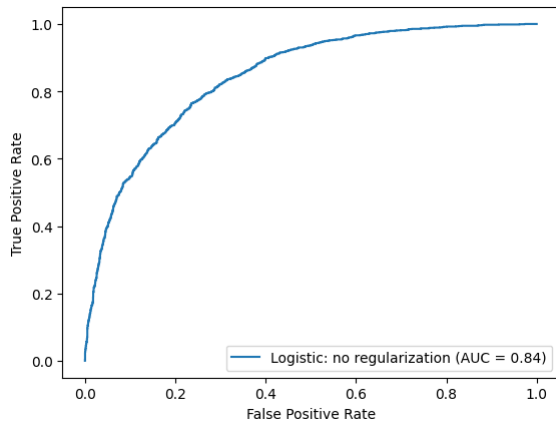


Figure 7. ROC curve with no regularization

Coefficient		Coefficient	
gender	0.0366	OnlineBackup	-0.061
SeniorCitizen	0.213	DeviceProtection	-0.0167
Partner	0.0956	TechSupport	-0.451
Dependents	-0.215	StreamingTV	0.278
tenure	-0.0345	StreamingMovies	0.394
PhoneService	-0.763	Contract	-0.631
MultipleLines	0.246	PaperlessBilling	0.343
InternetService	0.98	PaymentMethod	-0.171
OnlineSecurity	-0.301	MonthlyCharges	-0.000741
Intercept	-0.699		

Table 8. logistic regression (§3.1) model parameters