Machine Learning PSET 5

Weigeng Li G44406413

Problem 1

1.

$$P(Y=-)=\frac{9}{21}=\frac{3}{7};\ P(Y=+)=\frac{12}{21}=\frac{4}{7}$$

$$H(Y)=\ -\sum_{y\in Y}p(y)\log_2 p(y)=-\frac{3}{7}*\log_2\frac{3}{7}-\frac{4}{7}*\log_2\frac{4}{7}=0.985$$

2.

$$H(Y|X=x)=-\sum_{y\in Y}p(y|x)\log_2 p(y|x)$$

$P(Y=y\|X_1=x)$	$X_1=T$	$X_1=F$
$y=+$	7/8	5/13
$y=-$	1/8	8/13

$$H(Y|X_1=T)=-\frac{7}{8}*\log_2\frac{7}{8}-\frac{1}{8}\log_2\frac{1}{8}=0.5436$$

$$H(Y|X_1=F)=-\frac{5}{13}*\log_2\frac{5}{13}-\frac{8}{13}\log_2\frac{8}{13}=0.9612$$

Same Process as above we can get.

$P(Y=y\|X_2=x)$	$X_2=T$	$X_2=F$
$y=+$	7/10	5/11
$y=-$	3/10	6/11
$P(Y\|X_2=x)$	0.8813	0.9940

The distribution of X are

	T	F
$P(X_1=x)$	8/21	13/21
$P(X_2=x)$	10/21	11/21

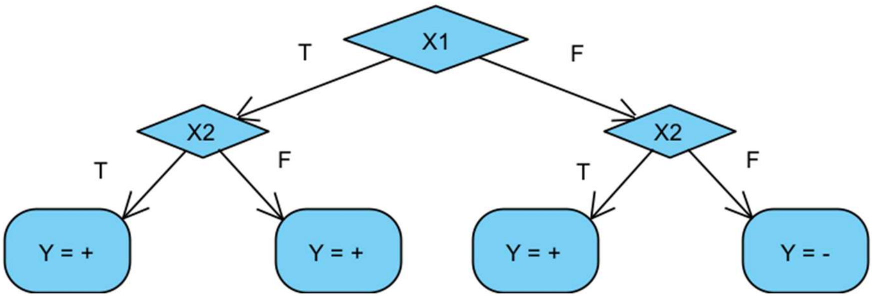
$$H(Y|X)=\sum_{x\in X}p(x)H(Y|X=x)$$

$$H(Y|X_1)=0.8021\ ;\ H(Y|X_2)=0.9403$$

$$IG(X_1)\equiv H(Y)-H(Y|X_1)=0.985-0.8021=0.1829$$

$$IG(X_2)\equiv H(Y)-H(Y|X_2)=0.985-0.9403=\ 0.0447$$

3. From the first section, the first node is X_1 because it has the highest information gain. The second node is X_2.



Problem 2

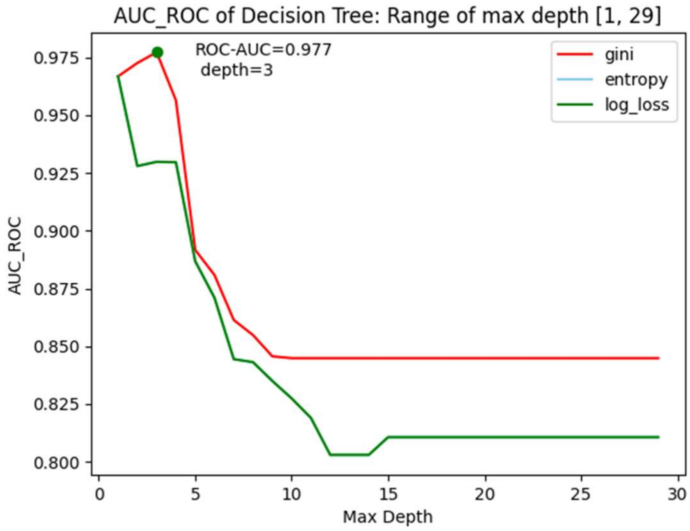
Preprocessing

This data set contains discrete and continuous variables. I scale all continuous variables into [-1, 1]. Discrete variables are 'Type1' and 'Type2'. They are transformed into One-Hot code which involves creating a binary column for each unique value in a categorical variable. For an example, 'Type1=Grass' transformed to a new variable 'Type1_Grass = 1'. This method is useful because the categories have no inherent order.

Method of Evaluation

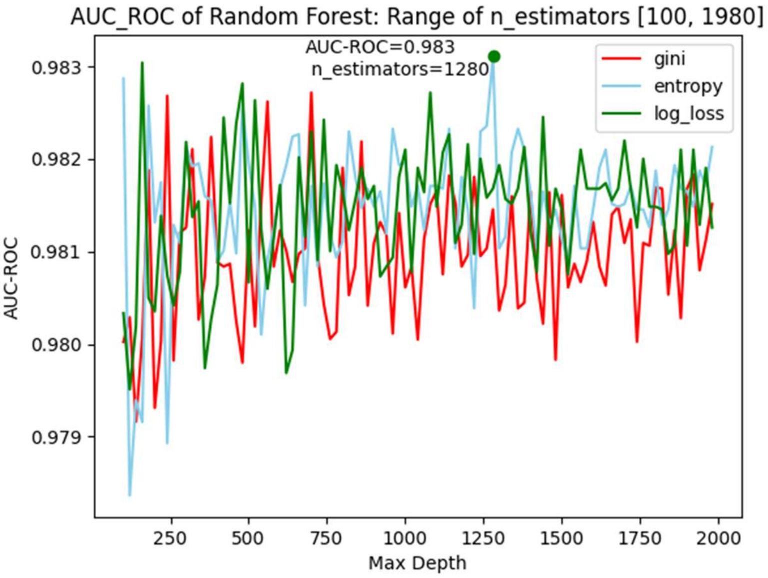
This is a classification problem with 65 positive (Legendary = True) sample and 735 negative sample. This means samples in this dataset is not balanced. If a model predicts all sample into negative, it will have $\frac{735}{800} = 91.875\%$ of accuracy. Therefore, Area Under ROC Curve (AUC-ROC) is a better index for evaluation. The ROC curve plots the true positive rate against the false positive rate for different threshold values, and the AUC-ROC measures the area under this curve. The AUC-ROC value ranges between 0 and 1, where 0.5 represents random guessing, and 1 represents a perfect model. I'm using cross validation score to get average AUC-ROC score.

Model Performance



This figure is the AUC-ROC of decision tree model in different max depth and criterion (function to measure the quality of a split). The decision tree reaches the highest AUC-ROC at depth=3. Before that point, the model is underfitting since there are not enough nodes in the tree to discriminate samples. After that model is over fitting, that fit too closely to the training data and become less generalizable. The importance of features are.

Feature	Importance
Total	0.9749
Sp. Atk	0.0251
Other	0.0



Random Forest have higher over-all AUC-ROC than decision tree. The highest AUC-ROC is 0.983 when it has 1280 decision tree as base estimator.

Feature	Importance
Total	0.2627
Sp. Atk	0.1289
Speed	0.0971
HP	0.0884
Sp. Def	0.0849
Attack	0.0830
Defense	0.0686
Generation	0.0397
Type 1_Psychic	0.0157
Type 1_Dragon	0.0145

Compared to decision tree, Random Forest consider more variables. Total and Sp. Atk still are top 2 importance variable.

In conclusion, **the best indicator in tree model is Total.** Random Forest have a construct a more complicate model with more variable in each tree. This helps Random Forest model have higher AUC-ROC score than decision tree.