Individual HW2: Predict Customer Churn (A0149810E)

For this assignment, I will be analysing a dataset from a telecommunication company and I will be creating machine learning model to predict whether a customer will churn based on the various features available. Predicting whether a customer will churn is important for the company to prevent further revenue losses and analysing the reason behind churning is a crucial step towards crafting strategies that will be able to minimise the rate of churning in the future.

During the data pre-processing stage, I discovered that the feature 'TotalCharges' includes some null values and I filled these null values using the formula TotalCharges = tenure * MonthlyCharges. I also converted all the categorical features into binary dummy variables, which will allow the inclusion of the categorical features in our machine learning model. Finally, I dropped the feature 'customerID' as it is a unique identifier that is randomly assigned to each customer and is unlikely to contribute to whether a customer will churn.

After data cleaning, I carried out exploratory data analysis by plotting graphs us to explore interesting relationship between the features and the label. Some features appear to be better predictor of whether a customer will churn because the proportion of customer possessing those features differ significantly between the churn customers and the non-churn customers. For example, I observed that customers will low tenure and high monthly charges are more likely to churn that those with high tenure and low monthly charges. This sounds reasonable as high monthly charges may cause customers to be dissatisfied with the company, causing them to leave the company prematurely as seen from the low tenure. Thus, 'MonthlyCharges' sounds like a potential important feature that should be used in our machine learning model.

In addition to the existing features in the dataset, I also observed some interesting pattern with regards to the number of services that a customer subscribes to. Features such as 'OnlineSecurity' and 'TechSupport' seem to be moderately correlated to 'Churn' and I wondered whether more services may lead to higher customer satisfaction, resulting in an inverse relationship between number of services and probability of churn. Hence, this leads me to engineer a new feature 'NumberOfSvcs' to find out how many internet services does a customer subscribe to.

For this dataset, which is a supervised classification problem, I have decided to use logistic regression and decision tree as the baseline models. I considered using k-

nearest neighbours but have decided that it may not be appropriate given the large number of categorical features present in the dataset. Linear regression model is also not an appropriate choice as the label we are predicting is binary, instead of continuous.

To further improve on these baseline models, I have also carried out ensemble learning methods, such as bagging, random forest and AdaBoost method. Ensemble learning methods combine several machine learning techniques into one model to reduce bias and variance, generally improving predictive ability of the model.

Before creating the model, we divided the data randomly into training and test dataset. However, as all our models work on this particular dataset, there is a chance of overfitting: the model fits this particular training set too well and hence sacrificing the accuracy when it comes to the test set. To reduce this risk of overfitting, I carried out k-fold cross validation, which splits the data into k sets and creates k different machine learning models based on k different combinations of training and test datasets. This reduces the reliance of the model on one particular training set and minimises overfit.

In my case, the logistic regression model yields the highest average cross validation score of 80.4%, followed by the AdaBoost model with average cross validation score of 80.2%. The important features in predicting churn include: TotalCharges, MonthlyCharges, tenure, NumberOfSvcs and Contract. This information would help in formulating strategies to minimise the rate of churn in the future. For example, as we know that more services lead to a lower probability of churn, the company may want to be more aggressive in promoting its different value-added services to customers that have subscribed to the internet service.

This model could be improved with the provision of additional features, such as customer income level, customer service ratings and address. The income level of a customer gives us a baseline against which the monthly charges can be compared. While a higher monthly charges may generally lead to higher probability of churn, customers of different income level may have different tolerance to the charges. Customer service ratings will tell us how satisfied is the customer with the company, which is relevant as customers who are dissatisfied by the company’s service tend to churn. Address may also be relevant as the network coverage of the telecommunication company may be weaker in certain areas, causing customers living there to have a higher probability of churn.