BT2101 HW2: CUSTOMER CHURN

MATRIC NO: A0168820B

NATIONAL UNIVERSITY OF SINGAPORE

For this assignment, I have put together a comprehensive analysis which includes new concepts from outside class and novel visualizations. I have ended off with a strategic recommendation for the telco company.

Overview

In this section, I printed out the shape, column names and data types to get a feel for the dataset. I also printed the counts of how many blank cells and NA values there are. At this point, my preliminary intuition was that *TotalCharges* might be superfluous as *Tenure* and *MonthlyCharges* are indicative of the same information.

Data Manipulation

I first discarded the blank values from the *TotalCharges* column by removing the rows. This is acceptable as only 11 out of 7032 rows were affected – a small percentage. I avoided filling in with **mean** or **mode** data since *TotalCharges* has a large range and the variance would be high; filling in proxy data might 'dilute' the model later. Then, I converted the *TotalCharges* column from string to numerical form. I also replaced the 'No Phone Service' and 'No Internet Service' values in several columns with just 'No', to keep the data as simple as possible and maximize **parsimony** (keeping in mind that the data has to be binarized later). Lastly, I **binned** the *Tenure* into bins of 12 months, in preparation to visualize by year during EDA.

Exploratory Data Analysis

First, I visualized the overall percentage of *Churn* customers. From this, we can tell that a blanket assignment of any new data point to 'No Churn' will be accurate 73.4% of the time. Our model will have to do better. I then plotted the frequency of each categorical variable against *Churn* to visualise how much that feature affects churn:

1. Gender has little to no effect on churn
2. Having a 1-or-2-year contract made churn less likely (quite intuitive)
3. Customers who pay by electronic cheque or who did not have a partner had more churn
4. The longer a customer's tenure (how long they have been a customer), the less likely they are to leave

I also visualised each numerical variable against *Churn*. Some insights that I gleaned were:

1. Having a low *MonthlyCharge* was indicative of lower churn and vice versa
2. Once a customer crosses the 1.5-year tenure barrier, his churn likelihood drops drastically

I noted that these insights indicate **correlation** but not necessarily **causation**. Also, to confirm my earlier intuition that *TotalCharges* might be correlated with *Tenure* and *MonthlyCharges*, I plotted charts and this turned out to be true. In model building, it would make sense to thus drop the *TotalCharges* column.

Data Processing

This section was primarily to prepare the data to be fed into the ML models. Binary variables were **encoded** into 1s and 0s, and other categorical variables (with 3 or more categories) were **binarized** into dummy variables. Numerical variables were **standardized** using standard scaler. I chose this method over min-max scaling as I observed that the numerical features were normally distributed when plotted against churn (from violin plots).

Model Building:

I performed chi2 test to compute the significance of each feature. The null hypothesis is that the feature has no effect on churn. Then we calculate the p-value, and if lower than 0.05, then we **reject** the null hypothesis with 95% confidence. *Gender* and *PhoneService* had p-values higher than 0.05 and thus were insignificant features.

For all models, I did a 70-30 train-test split. **Oversampling** was used for Decision Tree, Random Forest and Bagging since these models use **majority voting** in their algorithm, so it made sense to balance out the churn and non-churn data points to prevent **bias**. To avoid **information leakage**, I did the split before conducting oversampling, so the datapoints would not be repeated in the test set. For feature selection, I used a combination of the chi2 scores, the fact that *TotalCharges* is highly correlated with other features, and a bit of intuition. I used **trial and error** to test out a vast combination of feature selections for each model, and settled for a maximum balance between **parsimony** (to prevent overfitting) and **model accuracy**. The feature selection I performed was as such:

- Decision Tree – Selected top 3 categorical features (by chi2 score) and *Tenure* and *MonthlyCharges*. I limited this to 5 features to prevent **overfitting**.
- Random Forest – I dropped *PhoneService*, *Gender* and *TotalCharges* and took all the other features.
- Log Regression – I dropped *PhoneService*, *Gender* and *TotalCharges* and took all the other features.
- KNN – I dropped *PhoneService*, *Gender* and *TotalCharges* and took all the other features.
- Adaboost – I dropped *TotalCharges* only. Trial and error showed that the accuracy was much higher when all other variables were included, hence I decided to leave them in.
- Bagging – Chose the same features as in the decision tree. This is because the bagging algorithm chooses important **instances** and not **features**. Although the accuracy was slightly higher when I tried taking in all variables, the model was probably overfitted. I hence left it with just the top 5 features.
- SVM – I dropped *PhoneService*, *Gender* and *TotalCharges* and took all the other features.

Model Evaluation

I used both **k-fold cross-validation** and **test-set validation** on all models – the performance table is presented in the appendix. The **baseline accuracy** we established was 73.4% - all our models other than the decision tree surpassed this. Tree based models like DT, RF and Bagging didn't perform so well on this dataset, even with oversampling. SVM and Log Regression performed the best. In general, **False Negatives** were prevalent in a lot of the models – this is due to the disproportionately large number of 'No Churn' data points provided.

To reduce the False Negatives, I decided to **fine-tune** my hyperparameters by using the **gridsearch algorithm**. Although this did not lead to much gain in accuracy, my **recall (TP/P)** scores increased across the board, and the False Negatives decreased.

Additional Data and Further Discussion on the Business strategy

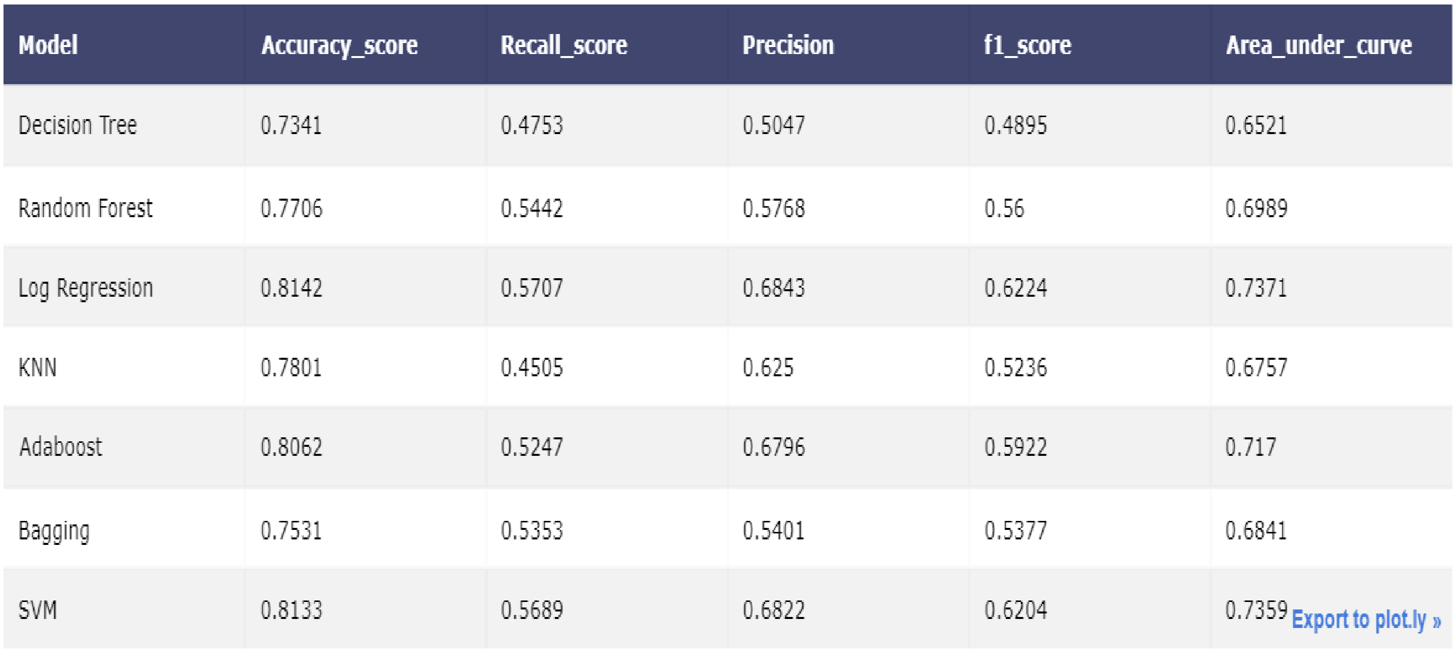
For this business problem, we seek to quickly identify the potential 'Churn' customers and launch customer retention tactics targeted at these individuals – the nature of this problem allows for False Positives (as we can write it off as a business cost), but False Negatives are penalised heavily. Looking at the business problem, if we simply wish to predict churn, why not just ask the customer if they are happy with the service, and what more would they like? Hence, additional data points include **customer satisfaction** and **service ratings**. Going forward, a good **strategy** for this company would be to offer tempting 2-year contract deals to 'lock in' customers at the start, and once they cross this barrier, their stickiness with us will significantly reduce churn likelihood. Another possible strategy would be to conduct **regular feedback surveys** on customers to gather the sentiment on the ground and identify potential churn in advance.

Appendix

Feature Importance Table

Feature	Chi-Square (ordered)	p-val >= 0.05
Phone Service	0.09	TRUE
Gender	0.25	TRUE
Multiple Lines	6.51	FALSE
Streaming Movies	15.93	FALSE
Streaming TV	17.32	FALSE
Device Protection	20.21	FALSE
OnlineBackup	31.21	FALSE
Payment Method Mailed Check	44.73	FALSE
Internet Service DSL	71.14	FALSE
PaymentMethod Bank Transfer (automatic)	76.62	FALSE
Partner	81.86	FALSE
Payment Method Credit Card	99.97	FALSE
Paperless Billing	104.98	FALSE
Dependents	131.27	FALSE
Senior Citizen	133.48	FALSE
Tech Support	135.44	FALSE
Online Security	147.17	FALSE
Contract One Year	176.61	FALSE
Internet Service No	285.48	FALSE
Internet Service Fiber Optic	372.08	FALSE
Payment Method Eletronic Check	424.11	FALSE
Contract Two Year	486.22	FALSE
Contract Month to Month	516.71	FALSE
Monthly Charges	3653.07	FALSE
Tenure	16377.32	FALSE
Total Charges	629630.81	FALSE

Model Performance Table



Model	Accuracy_score	Recall_score	Precision	f1_score	Area_under_curve
Decision Tree	0.7341	0.4753	0.5047	0.4895	0.6521
Random Forest	0.7706	0.5442	0.5768	0.56	0.6989
Log Regression	0.8142	0.5707	0.6843	0.6224	0.7371
KNN	0.7801	0.4505	0.625	0.5236	0.6757
Adaboost	0.8062	0.5247	0.6796	0.5922	0.717
Bagging	0.7531	0.5353	0.5401	0.5377	0.6841
SVM	0.8133	0.5689	0.6822	0.6204	0.7359

Model Performance Table After GridSearch

	Accuracy	Recall	Precision	f1 Score	AUC
Logistic Regression	0.7536	0.795	0.5241	0.6317	0.7668
Support Vector Machine	0.7408	0.8146	0.5078	0.6256	0.7643