SMU
SINGAPORE MANAGEMENT UNIVERSITY

Course: Data Mining and Business Analytics

AY2021-22 Term 2

Final Report:
Assessing the Risk of Telco Customer Churn using Predictive Models

Course Identification: IS424
(Section G1)

Authors:	Matriculation Number:
Pek Hui Leng Jolene	01379221
Wong Jia Yi	01417703
Lau Yi Ting, Glendys	01416345
Pan Zijiao	01446101
Ong Ai Wei	01417682

Content Page

1. Background and Information

Customer turnover is another term for customer attrition. Customer churn analysis and customer churn rate are frequently used as one of the company's key operational indicators by telephone operators, Internet service providers, insurance companies, and early warning monitoring service companies, because the cost of maintaining a customer is much lower than the cost of acquiring a new customer. Customer service departments are common in these businesses, and one of its responsibilities is to try to reclaim lost clients, because a loyal customer is worth far more than a new one in the long run. **Therefore, our group is interested to produce a customer churn model that, as far as possible, can accurately predict when a customer will churn, as well as produce insights into some reasons why they churn.**

Companies generally divide churn customers into two categories: active churn and passive churn. Active churn customers refer to those customers who decide to switch to other companies or use other services; passive churn customers are mostly due to changes in the basic environment, death or moving. Often, passive churn customers are not included in the analytical model. Churn analysis tends to focus on the analysis of active churn customers, because they churn due to more controllable factors such as if prices are becoming too high for them. These churn causes are generally related to the company-customer relationship over which the company controls to some extent the customer. For example, how to control bills or how to continue after-sales services.

To assess the customer's risk of churn, a customer churn model can be used to conduct a customer churn analysis. The customer churn model can prioritise possible churn consumers, allowing it to properly monitor churn-prone client groups. Business understanding, data preparation (feature engineering), modelling

analysis, and model deployment are the four basic components of model creation. Following that, the work content of each module will be shown in full, in conjunction with the project's background.

2. Motivation

Subscription-based services are negatively affected by every customer who leaves. This is because these customers are at the centre of their revenue. Customer retention should then be at the core of any of such businesses (Campbell, 2021). Churn will then help us know when customers are likely to stop using these services (Dorard & Patel, 2021). Big Data is making inroads into the business world, and churn uses it to analyse consumer behaviour in order to forecast when they are likely to stop using a company's services.

When it comes to developing a churn model, the churn rate is crucial. The churn rate must be calculated and used in the churn algorithm. To calculate this rate, we will need to know the number of customers who ceased using a service during a certain time frame and the number of customers who are still using the service; we will divide the former by the latter to get our churn rate (KDnuggets, 2017). At first, only large corporations employed churn modelling (Dorard & Patel, 2021). However, many prediction services have been established using APIs, and these services have assisted businesses of all types to learn from churn modelling and thrive further.

It is relatively simple to carry out operations that aid in overall customer retention, yet individual customer retention is not only difficult but also costly (Dorard & Patel, 2021) - in terms of time and money spent by firms when interacting with each customer. This necessitates a solution that is not just inexpensive but also quick and efficient (Dorard & Patel, 2021). Based on their experience and likelihood of churning from a service, we may design personalised solutions for each consumer. This helps to create a competitive advantage for the companies as well since most of them offer the same services.

Customers behave in a variety of ways. A customer's behaviour influences whether or not they use or churn from a service. It is critical for organisations to comprehend consumer behaviour and develop a model that will aid in the identification of each client experience.

3. Literature Review

We have identified the following papers that had also attempted to predict customer churn models. This section summarises each paper and includes the advantages, disadvantages as well as the limitations of the methods used and models produced.

3.1 A Customer Churn Prediction Model in Telecom Industry Using Boosting

SCI 2014《IEEE Transactions on Industrial Informatics》

Customer churn prediction is a main feature of modern telecom communication CRM systems. This research studies customer churn prediction in the real-world **and puts forward the use of boosting to enhance a customer churn prediction model**. Instead of using boosting as a method to boost the accuracy of a given basis learner like what most research papers do, this paper separates customers into two clusters based on the weight assigned by the boosting algorithm. This helps to clearly identify a higher risk customer cluster. Logistic regression is used here as a base learner, and a churn prediction model is built on each cluster, respectively. The result is contrasted **with a single logistic regression** model. Further evaluation done by the researchers also reveals that **boosting** helps to separate churn data well. Hence, boosting is highly recommended for churn prediction analysis (Lu, Lin, Lu, & Zhang, 2014).

3.2 A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn

SCI 2016《IEEE Transactions on Industrial Informatics》

This paper has discovered that when handling the amount of big data available for the telecom industry, the existing churn prediction models do not work very well to handle this data. Plus, decision makers are always faced with inaccurate operations management. As such, this paper proposes using a new clustering algorithm **called semantic-driven subtractive clustering method (SDSCM)**. Results have shown that SDSCM has stronger clustering semantic strength than other clustering methods such as subtractive clustering method (SCM) and fuzzy c-means (FCM). Afterwhich, a parallel SDSCM algorithm is implemented through a Hadoop MapReduce framework. In the case study, the proposed parallel SDSCM algorithm enjoys a fast-running speed when compared with the other methods such as serial SDSCM, especially with datasets of between 100 to 200k. (W. Bi, M. Cai, M. Liu and G. Li, 2020). This was an interesting paper to read, after studying this paper, we realised that this could be something that we can work on in the future as well, to use a similar clustering method for a more accurate churn modelling prediction.

3.3 Evaluation of customer behaviour with temporal centrality metrics for churn prediction of prepaid contracts. SCI 2020《EXPERT SYSTEMS WITH APPLICATIONS》

Through this paper, we learned that churn analysis modelling is usually considered a classification problem and past data mining techniques used includes the following: decision-rule based classifier, decision tree approaches, neural networks, nearest neighbour, ensemble methods, logistic regression and support vector machine (SVM). As such, we have also tried to implement most of these in our project. This paper suggests that call detail records (CDRs) have since become a highly sought-after source of data for churn prediction, in the sense where the chances of a customer churning can be studied and derived from that of their churned friends. Many researchers have used binary classification methods to predict churn of customers. Through

utilising a few machine learning techniques, **some of them verify that customers' social relationships influence the decision of changing the operator. Therefore, this paper uses a novel method to extract the dynamic relevance of each customer using social network analysis techniques with a binary classification method called similarity forests.** The dynamic importance of each customer is determined by applying various centrality metrics over temporal graphs, to represent the relationships between customers and to extract behavioural patterns of churners and non-churners. These relationships are established in a temporal graph using the CDRs of the customers. They, then, proceed to compare the performance of different centrality metrics applied over two types of temporal graphs: Time-Order Graph and Aggregated Static Graph (Calzada-Infante, Laura & Óskarsdóttir, María & Baesens, Bart, 2020).

3.4 Study on the Prediction of Imbalanced Bank Customer Churn Based on Generative Adversarial Network. SCI 2020《Journal of Physics Conference》

Our dataset has an imbalance rate of 26.6%. With this number on hand, we set out to find a paper which dealt with imbalance data for classification methods, and especially for churn analysis or churn-modelling related papers. This paper talks about customer churn in the aspect of commercial banks, bypassing the few differences that are not important, it is rather similar to that of the telecom industry. In the context of this paper, the imbalanced commercial bank customer data will cause unpredictability to the minority class. Aside from common ways such as SMOTE, this paper proposes **improving imbalanced data using generative adversarial networks (GAN)** to deal with the issue of poor prediction performance of traditional classifiers for minority class. The classifier is then used to train the balanced data to improve the prediction performance of the minority class. For this experiment, the data of a typical commercial bank customer was measured with indicators such as F1, Precision, and compared **with traditional data sampling methods such as SMOTE, BSSMOTE**. It is discovered that using GAN is more feasible and applicable to the classification of imbalanced data of banks which has better application value (Li, Bo & Xie, Jiuzuo, 2020). We have also, thus, attempted to use GAN as part of evaluating our dataset (which we will further explain in later sections).

4. Dataset

After much research, we sized down to 3 datasets which are most popular for our problem topic. The first dataset **(Telco Customer Churn)** is from Kaggle[1]. There are 2 limitations to this dataset - it is rather small with 7043 rows (customers) and 21 columns (features), and it has an imbalance rate of 26.6%. The second dataset[2] has a similar number of records but is missing a higher number of dates. The third dataset[3] is provided by an individual which we felt was unreliable. Hence, we only used the first dataset as our project's

[1] https://www.kaggle.com/blastchar/telco-customer-churn/version/1
[2] https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113
[3] https://www.datafountain.cn/datasets/5667

main dataset. We tried to find a bigger dataset, however, such datasets are not readily available as open sources online. Therefore, we will be using certain analysis techniques to deal with its 2 limitations aforementioned. In this project, a variety of algorithms are used to build prediction models, and the recall rate, precision rate, AUC, and accuracy rate of various models are compared, and then the most suitable algorithm is selected and continuously optimised.

4.1 Variables Description

```
In [3]: pd.head()
```

Out[3]:

ineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	...	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
Yes	...	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
Yes	...	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
Yes	...	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	...	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

Figure 1 - First 5 rows of our dataset

Variable	Data type	Variable Content
customerID	str	Id
gender	str	Gender
SeniorCitizen	int	yes 1 no 0
Partner	str	If has partner yes no
Dependents	str	If has dependens yes no
tenure	int	Months of use of the service
PhoneService	str	If has phone service yes no
MultipleLines	str	If has multi-lines yes no
InternetService	str	DSL/fiber optic/no
OnlineSecurity	str	If has security service yes/no/no internet service
OnlineBackup	str	If has backup service yes/no/no internet service
DeviceProtection	str	If has device protection service yes/no/ no internet service
TechSupport	str	If has tech support yes/no/ no internet service
StreamingTV	str	yes/no/ no internet service
StreamingMovies	str	yes/no/ no internet service
Contract	str	Contract methods month-to-month/one year/two year
PaperlessBilling	str	If has paperless bill Yes/no
PaymentMethod	str	Electronic check/mailed check/bank transfer/automatic/credit card(automatic)

MonthlyCharges	float	Charges per month
TotalCharges	float	Total charges until now
Churn	str	If churn Yes/no

Table 1 - Variable Description

From the sorted information in Table 1 and Figure 1, the size of the dataset is (7043, 21), and no missing values are used in each type of data. The **[Churn]** column is the label result and the target column: whether to churn. The dataset contains labelled results, and the model is suitable for building with supervised learning algorithms.

Combined with the relevant knowledge of the operator's industry, whether the user is lost may be related to the operator's service content, the user's own situation, usage habits and other factors. For example, whether the user's economy is independent is very likely to affect whether the user will churn. The following is to explore the information of each column of data one by one, which will be explained further in **Section 9: Exploratory Data Analysis.**

5. Data Preprocessing

```
# change value to float
data['TotalCharges'] = data['TotalCharges'].astype(float)

# change'No internet service' to 'No'。
replace_cols = ['OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
                'TechSupport', 'StreamingTV', 'StreamingMovies']
for i in replace_cols:
    data[i] = data[i].replace({'No internet service': 'No'})

# replace str
data['SeniorCitizen'] = data['SeniorCitizen'].replace({1:'Yes', 0:'No'})

# change Tenure value to range
def tenure_lab(data):
    if data['tenure'] <= 12:
        return 'Tenure_0_12'
    elif (data['tenure'] > 12) & (data['tenure'] <= 24):
        return 'Tenure_12_24'
    elif (data['tenure'] > 24) & (data['tenure'] <= 48):
        return 'Tenure_24_48'
    elif (data['tenure'] > 48) & (data['tenure'] <= 60):
        return 'Tenure_48_60'
    elif data['tenure'] > 60:
        return 'Tenure_gt_60'
data['tenure_group'] = data.apply(lambda data: tenure_lab(data), axis=1)

# seperate customer into churn or not churn group
churn = data[data['Churn'] == 'Yes']
not_churn = data[data['Churn'] == 'No']

# Separate categorical variables from numeric variables
Id_col = ['customerID']
target_col = ['Churn']
cat_cols = data.nunique()[data.nunique()<6].keys().tolist()
cat_cols = [x for x in cat_cols if x not in target_col]
num_cols = [x for x in data.columns if x not in cat_cols+target_col+Id_col]
# binary
bin_cols = data.nunique()[data.nunique()==2].keys().tolist()
# multityp
multi_cols = [i for i in cat_cols if i not in bin_cols]
```

Figure 2

We noted the attribute **`TotalCharges`** has some null values hence we removed the records with the null values (as the missing values only account for 0.015%). We also replaced some columns with binary values instead of their original **`'Yes'`** and **`'No'`** values to cater to our machine learning model. Additionally, as there is a wide range of values when it comes to **`tenure`**, hence we decided to break them into 5 main ranges of values which transformed them into categorical attributes. Finally, we sorted the customers into 'churn' and 'non-churn' groups respectively and separated the categorical variables from the numeric variables for exploratory data analysis (EDA).

```
: # preprocessing
from sklearn.preprocessing import LabelEncoder, StandardScaler

Id_col = ['customerID']

cat_cols = data.nunique()[data.nunique()<6].keys().tolist()
cat_cols = [x for x in cat_cols if x not in target_col]

num_cols = [x for x in data.columns if x not in cat_cols+target_col+Id_col]

bin_cols = data.nunique()[data.nunique()==2].keys().tolist()

multi_cols = [i for i in cat_cols if i not in bin_cols]

# label encoding
le = LabelEncoder()
for i in bin_cols:
    data[i] = le.fit_transform(data[i])


data = pd.get_dummies(data=data, columns=multi_cols)

# standardization
std = StandardScaler()
scaled = std.fit_transform(data[num_cols])
scaled = pd.DataFrame(scaled, columns=num_cols)

# Delete the original numeric column and
# merge the standardized numeric column with the original data table
df_telcom_og = data.copy()
data = data.drop(columns=num_cols, axis=1)
data = data.merge(scaled, left_index=True, right_index=True, how='left')
```

Figure 3 - Encoding and Standardising of our attributes Section 4.1 of Jupyter Notebook

To better fit data into our machine learning model, we encoded our data to convert categorical variables into numerical values. Afterwhich, we carried out standardisation and replaced the original numerical columns with the new standardised numerical columns.

6. Metadata of our dataset

Metadata, simply put, is data about data. It is data used to represent other data. Metadata is mainly used to describe the properties of the data, which is of great significance for the organisation as it defines the structure of the data, as well as for future data analysis, visualisation, and modelling. It is important to us because in our project, we often encounter data with high dimensionality and many features. In order to facilitate further analysis, we must organise the data in a structured way, which metadata helps to do so. It enables us to understand our dataset more deeply. It also helps with our EDA and better feature and attribute selection to use for feature engineering.

meta_name	meta_role	meta_scale	meta_dtype	meta_remain	meta_unique
customerID	ID (not attributes)	nominal	object	False	n/a
gender	attributes	binary	object	Ture	2
SeniorCitizen	attributes	binary	object	Ture	2
Partner	attributes	binary	object	Ture	2
Dependents	attributes	binary	object	Ture	2
tenure	attributes	interval	int64	Ture	72
PhoneService	attributes	binary	object	Ture	2
MultipleLines	attributes	nominal	object	Ture	3
InternetService	attributes	nominal	object	Ture	3
OnlineSecurity	attributes	binary	object	Ture	2
OnlineBackup	attributes	binary	object	Ture	2
DeviceProtection	attributes	binary	object	Ture	2
TechSupport	attributes	binary	object	Ture	2
StreamingTV	attributes	binary	object	Ture	2
StreamingMovies	attributes	binary	object	Ture	2
Contract	attributes	nominal	object	Ture	3
PaperlessBilling	attributes	binary	object	Ture	2
PaymentMethod	attributes	nominal	object	Ture	4
MonthlyCharges	attributes	interval	float64	Ture	n/a
TotalCharges	attributes	interval	float64	Ture	n/a
Churn	target (churn or not)	binary	object	Ture	2
tenure_group	attributes	nominal	object	Ture	5

Figure 4 - Metadata of our data

In Figure 4, this metadata describes our data (listed as **`meta_name`**) more deeply in terms of:

- its role in our dataset (is it used as an attribute in our dataset or not? - e.g. **`customerID`** is not used in our dataset because its ID has no proper meaning and will not contribute to our analysis)
- its scale (nominal, binary, interval)
- its data type (object, integer, float)
- whether it remains as attributes for us to use in our analysis (e.g. since **`customerID`** is no longer considered as an attribute, it will not remain in our dataset)
- the number of unique values that describes the attributes (ranging from 2 to 72, **`'n/a'`** means we are unable to count the unique values for this attribute)

	meta_role	meta_scale	total
0	ID (not attributes)	nominal	1
1	attributes	binary	12
2	attributes	interval	3
3	attributes	nominal	5
4	target (churn or not)	binary	1

Figure 5 - Summary of our metadata analysis

From our analysis, we conclude that (using Figure 5) our metadata consists of 1 non-attribute column (which we will not be using). We noticed that we have 12 binary attributes, 3 interval attributes and 5 nominal attributes to work with. Finally, our **target** column (`Churn`) to work with contains binary values.

Hence, as demonstrated in Figures 4 and 5 above, our metadata provides information about one or more aspects of the dataset we have. It gives us a clear overview of what data we are working with and how we can best deal with this data. This will help us to develop more accurate models for our churn analysis.

7. Tools and Resources

Data Preparation	EDA	Modelling Analysis	Model Evaluation

• matplotlib
• seaborn
• plotly
  ‣ offline
  ‣ graph_objs
  ‣ figure_factory
	• numpy
	• pandas
	• scikit-learn
	  ‣ pre-processing
	  ‣ decomposition
	• imblearn
	  ‣ SVC SMOTE
	  ‣ KMeans SMOTE
		• sklearn.model_selection
		• sklearn.neighbors
		• sklearn.svm
		• sklearn.ensemble
		• sklearn.tree
		  ‣ Adaboost
		  ‣ GBDT
		  ‣ ExtraTree
		• xgboost
		• lightbgm
		• keras
		• Pytorch (Deep-learning)
			• sklearn.metrics

8. Customer Lifetime Value (LTV) Analysis

Customer lifetime value measures the total worth of a customer to a business over the whole period of their relationship, in this case, the financial worth of each customer to the Telco company. It is an important metric to analyse churn rate and develop strategies.

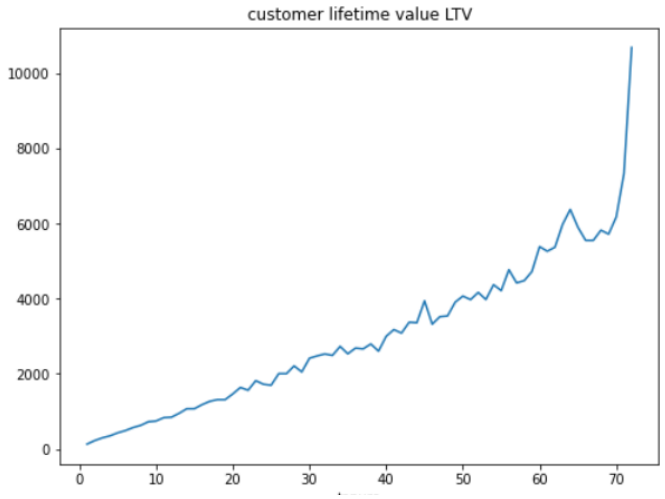

Figure 6 - Customer LTV, historical payment against remaining charges

Therefore, to calculate the customer lifetime value, we *combined total historical payment (total charges) and remaining charges* and plotted it against tenure. Figure 6 proves that the longer the tenure, the higher the customer lifetime value.

9. Exploratory Data Analysis (EDA)

9.1.1 Nominal Attributes

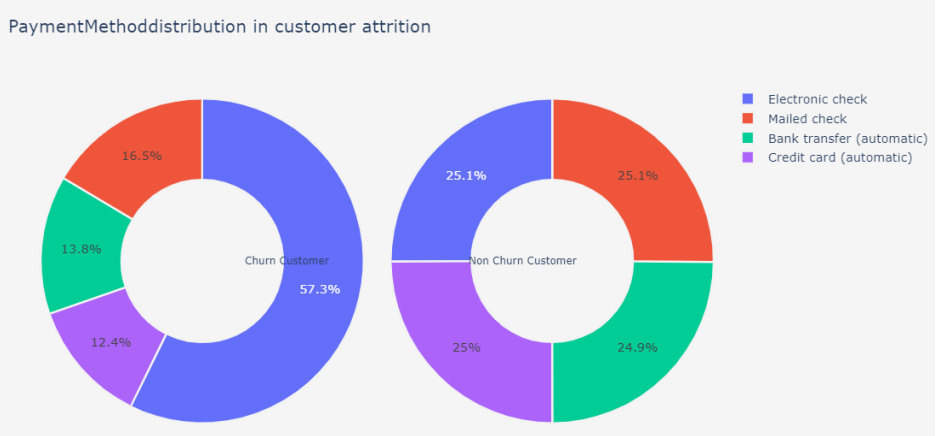

Figure 7 - Pie Chart on PaymentMethod Variable

Based on our original dataset **(refer to Table 1 and Figure 1)**, we also did exploratory data analysis by visualising the target class proportion of each individual attribute. For nominal attributes, we constructed a pie chart. In figure 6, we realised that for customers who did not churn, they are equally separated within the 4 payment methods. Whereas for the churn customers, surprisingly 57.3% of them paid by electronic check. This is an unexpected factor which requires further investigation.

9.1.2 Numeric Attributes

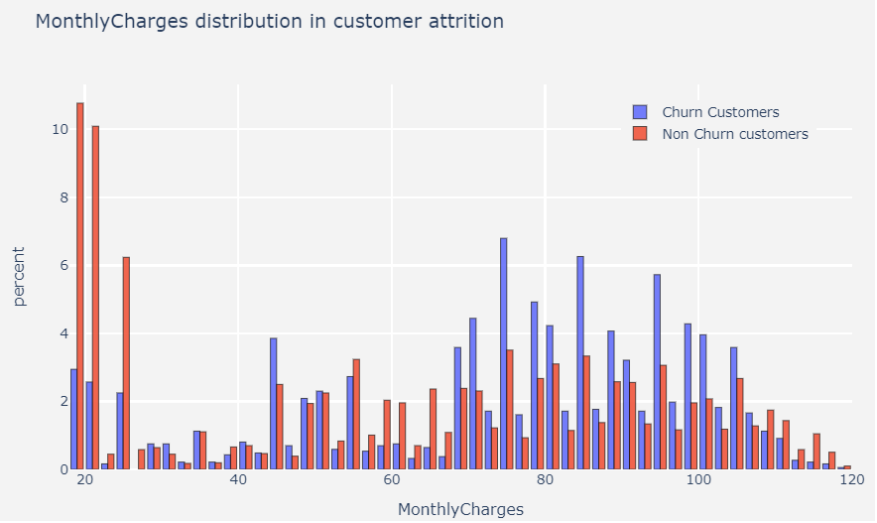

Figure 8 - Histogram on MonthlyCharges Attribute

As for numeric variables, we plotted histograms to understand each attribute relationship with the target classes. In figure 7, we focused on the monthly charges attribute. We can tell that as monthly charge increases, the percentage of no churn customers is concentrated around $0 - $28. Whereas the percentage of customers who churned is concentrated around $72-$130 which is higher than the non-churn customers, implying that the high monthly bill amount strongly contributes to the churn rate.

9.1.3 Customer churn grouped by tenure variable

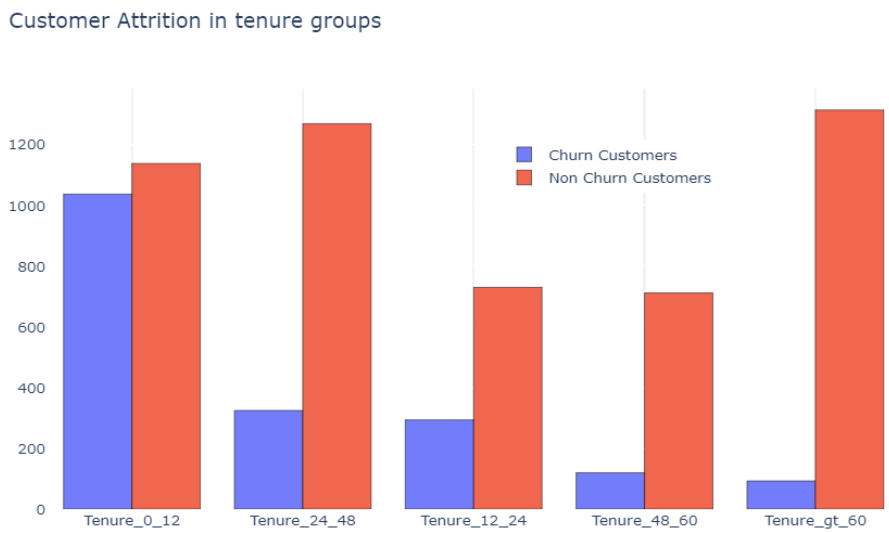

Figure 9 - Histogram of Tenure Variable

In figure 8, we are focusing on the Tenure variable (length of contract), we can see that as the tenure length increases, the number of churn customers decreases while the number of churn customers peaked at tenure length of more than 60. This shows that the longer the customer stays with the company, the smaller the probability of them churning.

9.1.4 Average spend grouped by tenure

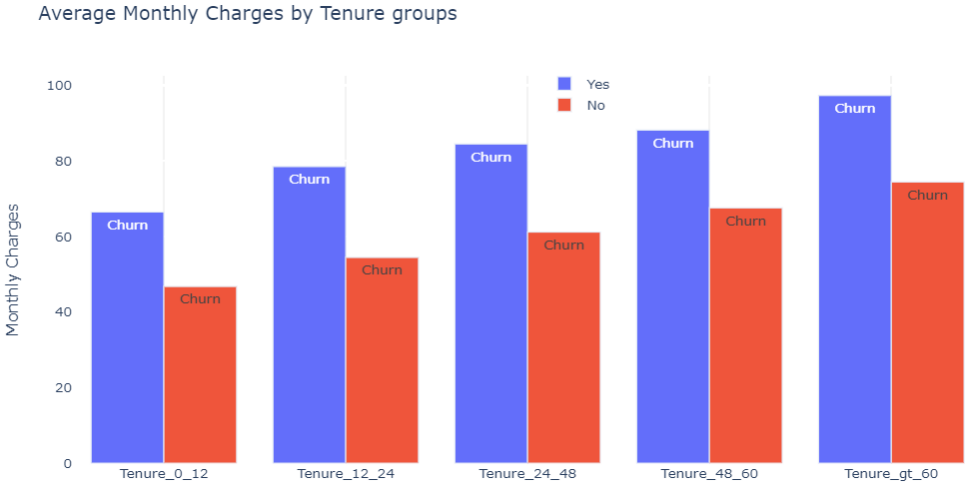

Figure 10 - Histogram of Average Monthly Charges against Tenure groups

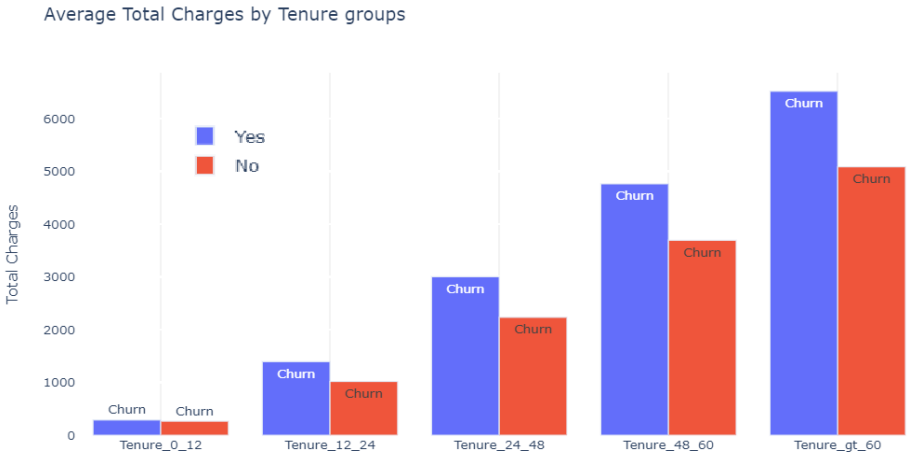

Figure 11 - Histogram of Average Total Charges against Tenure groups

Another important factor in our analysis is the total charges as it refers to the revenue earned by the Telecom company. And thus, to analyse this, we averaged expenditure and plotted it against tenure groups. Regardless whether customers churn or not, the average monthly consumption increases steadily as the duration of their stay increases. However, one thing to take note is that customers who churn have a higher average expenditure which could mean that the charges are a huge factor in one's retention.

9.1.5 Visualisation of customer persona

To understand customers' profile better, we constructed a customer persona. They allow us to understand key traits within them which can allow the Telco company to deliver a more personalised experience to increase retain rate. And to do so, we used radar charts. A radar chart displays multivariate data in a form of 2-dimension chart of quantitative variables represented on axes originating from the centre. We decided to use this as it allows us to compare the different entities with our large number of variables. Radar graph is also able to identify to what proportion the customers exhibit the attribute.

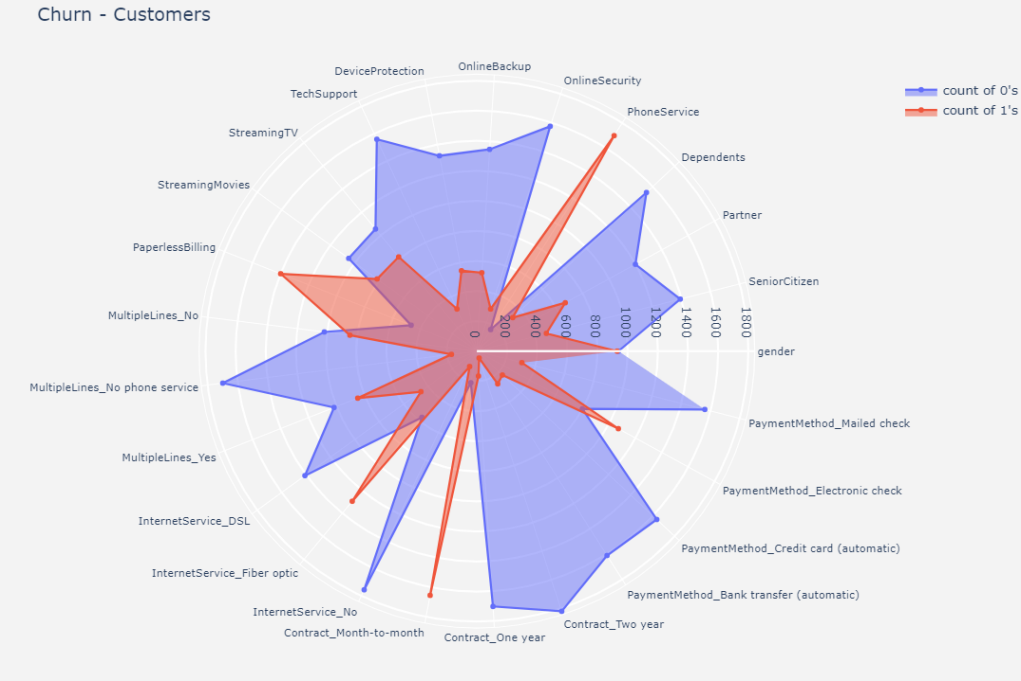

Figure 12 - Radar Graph of Churn Customers Attributes

For customers who churned, we can clearly see that they subscribe to phone service, do not subscribe to internet service, have not requested for online security, their contracts are on a monthly basis and opt for paperless billing. However the attribute of subscribing to phone service is not useful to us as it is the basic service provided by the Telco company. Additional investigation is needed for the paperless billing factor.

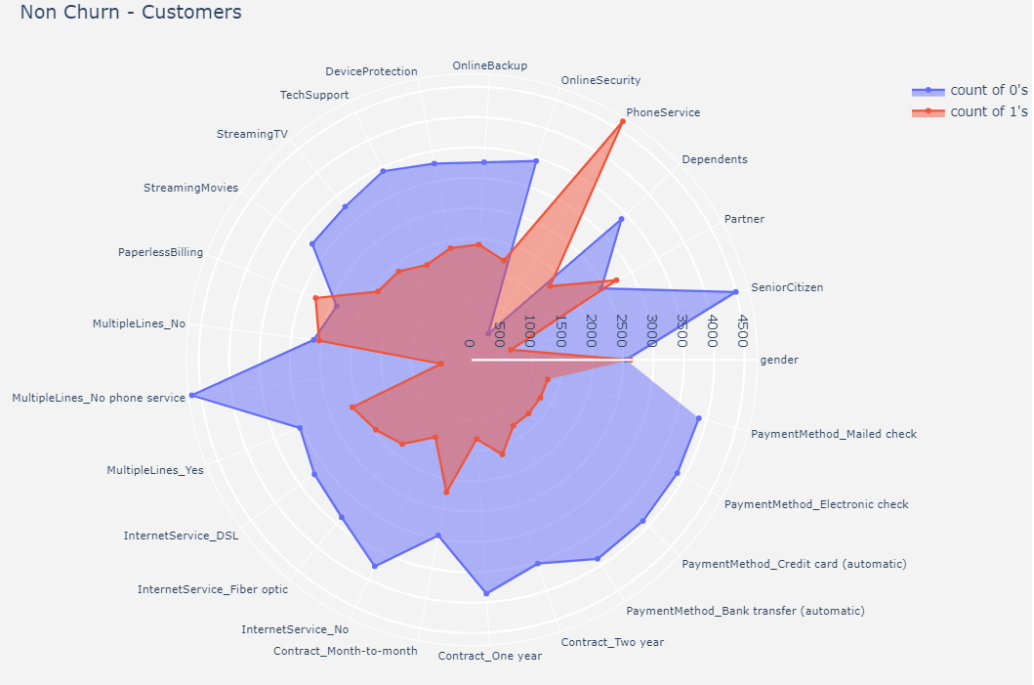

Figure 13 - Radar Graph of Non-churn Customers Attributes

For customers who churned, they do not subscribe to multiple lines of phone service and are not senior citizens. This shows that if a customer business with the Telco involves multiple people (such as their partner), they will be more loyal towards the Telco service.

9.2 Imbalance data

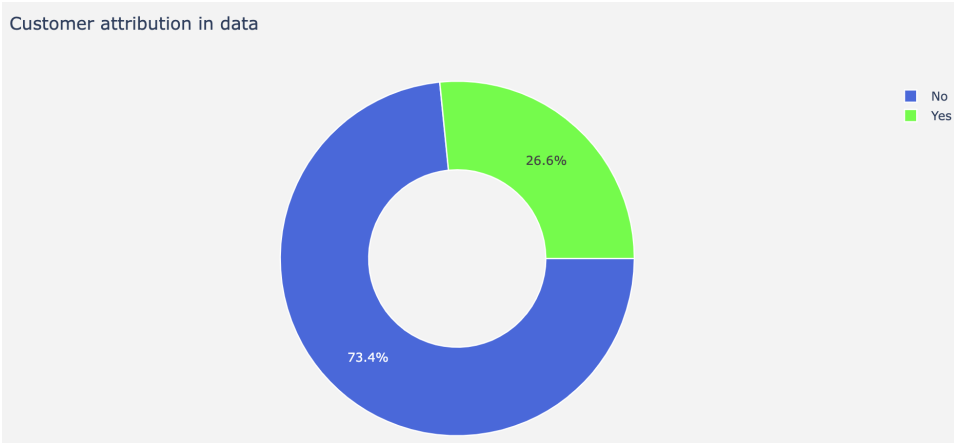

Figure 14 - Pie chart of our data proportion of each target class

Our dataset is slightly imbalanced with 73.4% of customers being churned customers and 26.6% being non-churned customers. However, we know that for any prediction models, using a slightly imbalanced dataset together with our smaller dataset, can affect the accuracy of any models generated. And thus, from our research and literature review, we decided to implement oversampling methods such as SMOTE and GAN.

10. Retention Forecast

10.1.1 SMOTE

SMOTE generates minority classes by taking a sample of each minority class and giving a synthetic example along the line segment connecting the nearest neighbour of all selected N minority classes. According to the required sample size, randomly select closest samples from N nearest neighbours. This process is described in SMOTE as shown in figure 1 below. First, for each observation x of the minority class, its N nearest neighbours are identified, as shown in the square sample in the figure. Then randomly select N neighbours. Finally, minority class sample is copied along the line that connects the real sample n with its nearby samples (Li, Bo & Xie, Jiuzuo, 2020).

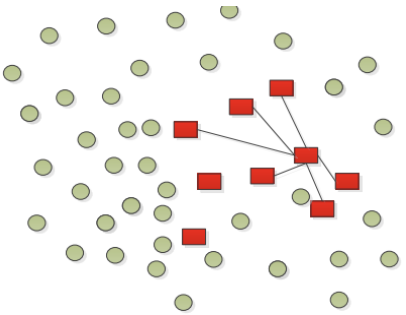

Figure 15 - Smote algorithm diagram (square is minority class, round is majority class)

10.1.2 GAN

The Generative Adversarial Network is a kind of network structure which can obtain new samples by using training samples. GAN's main task is to estimate the sample distribution of the training set, and then use the sample distribution to generate another sample similar to the training set (Li, Bo & Xie, Jiuzuo, 2020). Figure 16 below is a flowchart:

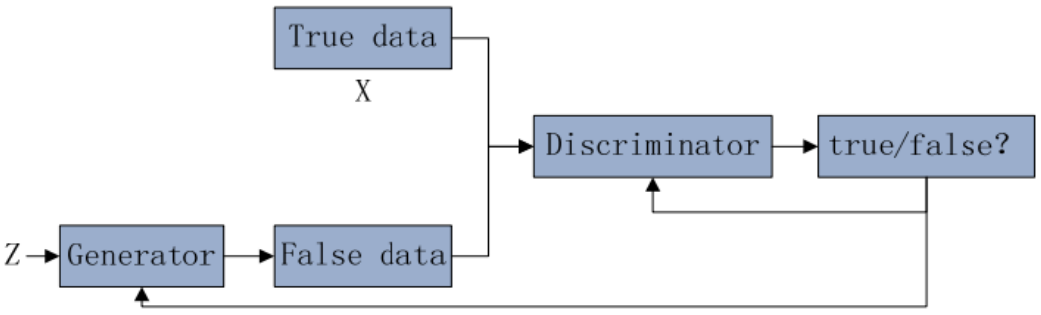

Figure 16 - The GAN Chart

At present, this model is widely used in image vision, anomaly detection and Credit card fraud. Compared with the traditional generation model, the GAN model doesn't need to generate synthetic data based on real data to approximate real data.

10.1.3 SMOTE or GAN

In order to determine which of the oversampling methods will be more effective in our project and forecast, we generated a random forest tree using the train and test dataset oversampled using SMOTE and GAN separately.

For SMOTE, accuracy and recall that are derived from the confusion matrix are 0.76109 and 0.77551 respectively, while the AUC is 0.85 (Figure 16) . For GAN, accuracy and recall that are derived from the confusion matrix are 0.75882 and 0.77551 respectively, while the AUC is 0.85 (Figure 17).

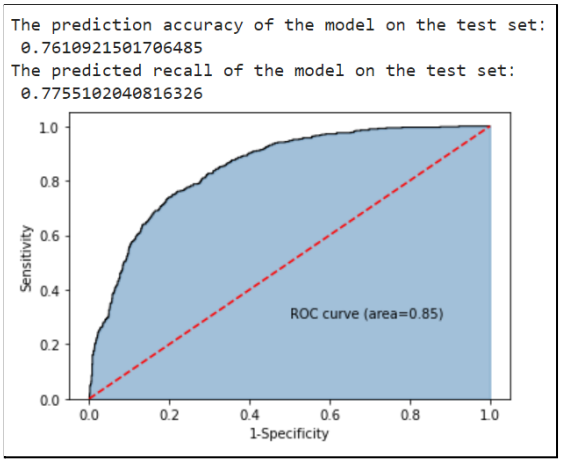

Figure 17: Accuracy, Recall, Confusion Matrix and AUC of Random Forest for SMOTE

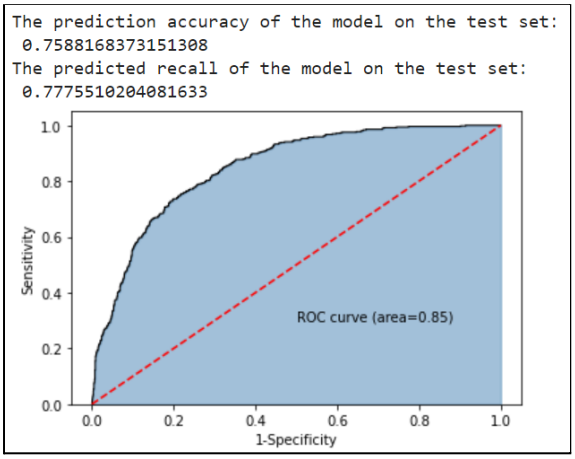

Figure 18: Accuracy, Recall, Confusion Matrix and AUC of Random Forest for GAN

Since SMOTE and GAN's recall and AUC are equal, we compare the accuracy instead. SMOTE's accuracy is higher, hence we went ahead with using the Training and Test set generated by SMOTE for the rest of the 11 models. SMOTE is better than GAN as lost users in the sample only accounts for 25% of the sample and GAN works better with even more imbalanced datasets (e.g. 5% imbalanced rate) because compared with SMOTE, the GAN model does not need to generate synthetic data based on real data to approximate the real data (Li, Bo & Xie, Jiuzuo, 2020). Thus, we used the SMOTE algorithm to balance and improve the data set before creating any models.

10.2 Generating a Random Forest Tree Without Oversampling

We also generated a Random Forest Tree using the raw dataset in order to evaluate the effectiveness before continuing generating the 11 models.

```
The prediction accuracy of the model on the test set:
 0.7952218430034129
The predicted recall of the model on the test set:
 0.45714285714285713
```

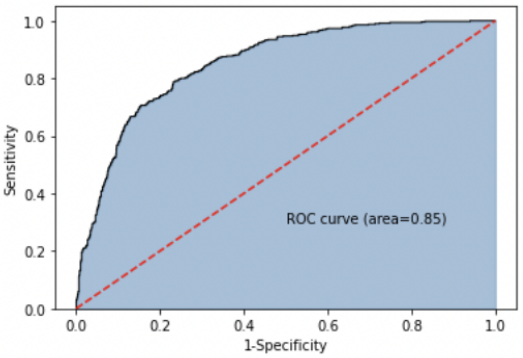

Figure 19: Accuracy, Recall and AUC of Random Forest **(raw) without dealing with imbalance data**

Generating a random forest tree using the raw dataset has resulted in a prediction accuracy of 0.80, recall of 0.46 and lastly, area under curve of 0.85. As previously mentioned, we will have a higher emphasis on recall as a metric and in this case, it is obvious that the recall rate is significantly lower than the random forest tree generated with the oversampled data be it SMOTE (0.77) or GAN (0.77).

This proves that using an oversampling method will indeed increase the accuracy of our prediction methods.

11. Model Performance

11.1 Grid Search and Metrics Measured for Models

For all the following 11 models, grid search was used for hyperparameter tuning to determine the optimal values for a given model. The hyperparameters are configured up-front and are provided by the caller of the model before the model is trained. Eventually, the selected parameters are the ones that maximise the accuracy score as the parameter search uses the score function (accuracy being the default) of the estimator to evaluate a parameter setting.

We then fit the models to predict churn of customers. Metrics such as accuracy, recall, ROC curve to find AUC and confusion matrix of the model on the test set for each model are calculated and compared against one another to select the best performing model(s). The higher the score of the metrics, the better the model is in assessing the risk of telco customer churn.

Model optimization goal: Considering that the problem focuses on the churn rate, we should try to find all the lost users and formulate a policy to retain as many customers as possible. Therefore, the recall rate of the model should be focused on when predicting.

Below is an example of an AUC-ROC curve and confusion matrix of random forest. These graphs were generated for all 11 models.

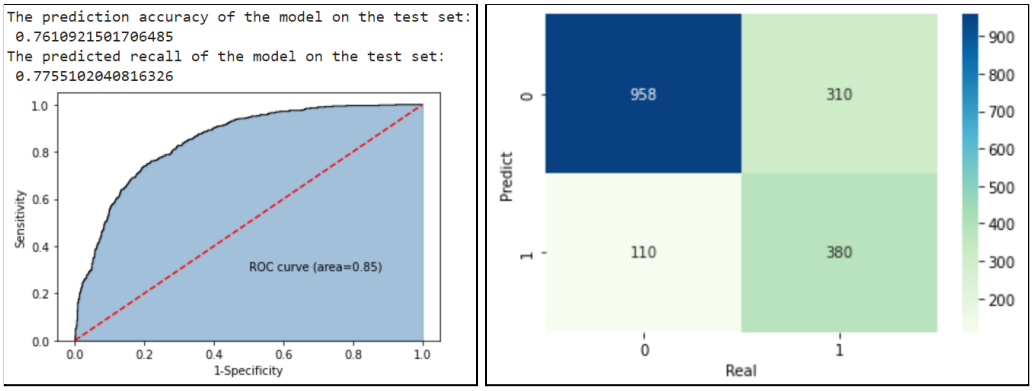

Figure 20: Accuracy, Recall, Confusion Matrix and AUC of Random Forest

11.2 Comparison and Recommendation of Models

The 11 models used are categorised into single model and fusion model. For single model, Random Forest, Decision Tree, Adaboost, K-Nearest Neighbour (KNN), Support Vector Machine (SVM) Linear and Nonlinear, LightGBM, XGBoost, Extra Tree and Gradient Boost Decision Tree (GBDT) and Deep Learning Artificial Neural Network (ANN) were used. For the fusion model, Voting was used.

A comparison table to compare the recall, accuracy and AUC scores across all 11 models is shown below. The recommended models, Adaboost, SVM Linear and Voting, have the highest AUC of 0.85. Out of the 8 models with 0.85 AUC score, Adaboost, SVM Linear and Voting have the top 3 recall scores of 0.78571, 0.80816 and 0.83878 respectively, and hence are chosen (highlighted in green).

Model Name	**Best Parameter**	**Recall**	**Accuracy**	**AUC Scores**
Random Forest	{'n_estimators': [700], 'max_depth': [8], ‘min_samples_split':[10], 'min_samples_leaf':[20]}	0.77551	0.76109	0.85
Decision Tree	{'max_depth': 8}	0.76327	0.72923	0.82
Adaboost	{'n_estimators': 600}	0.78571	0.76962	0.85
KNN	{'n_neighbors': 100}	0.85918	0.71388	0.84
SVM Linear	{'C': 0.8, 'gamma': 1.0 ,'kernel': 'linear'}	0.80816	0.75085	0.85
SVM Nonlinear	{'C': 0.8, 'gamma': 1.0 ,'kernel': rbf}	0.76222	0.51020	0.78
LightGBM	{'n_estimators': 100}	0.72653	0.77019	0.85
XGBoost	{'n_estimators': 100}	0.78163	0.74118	0.85
Extra Tree	{'max_depth': 8, 'min_samples_leaf': 5,	0.77551	0.76337	0.85

	'min_samples_split': 10, 'n_estimators': 600}			
GBDT	{'max_depth': 5, 'min_samples_leaf': 10, 'min_samples_split': 20, 'n_estimators': 100}	0.76122	0.75119	0.85
Deep Learning ANN	{'loss': 'binary_crossentropy', 'optimizer': 'SGD()', 'metrics': 'accuracy'}	0.73673	0.77418	0.84
Voting	Dependent on the previous models	0.83878	0.72526	0.85

Table 2: Comparison Table of 11 Models

11.3 Testing with Single Model

11.3.1 Random Forest¶

Upon using grid search, it was found that Random Forest's best parameter is {'n_estimators': [700], 'max_depth': [8], 'min_samples_split':[10], 'min_samples_leaf':[20]} that results in the best accuracy score of 0.803338.

The accuracy and recall that are derived from the confusion matrix are 0.76109 and 0.77551 respectively, while the AUC is 0.85.

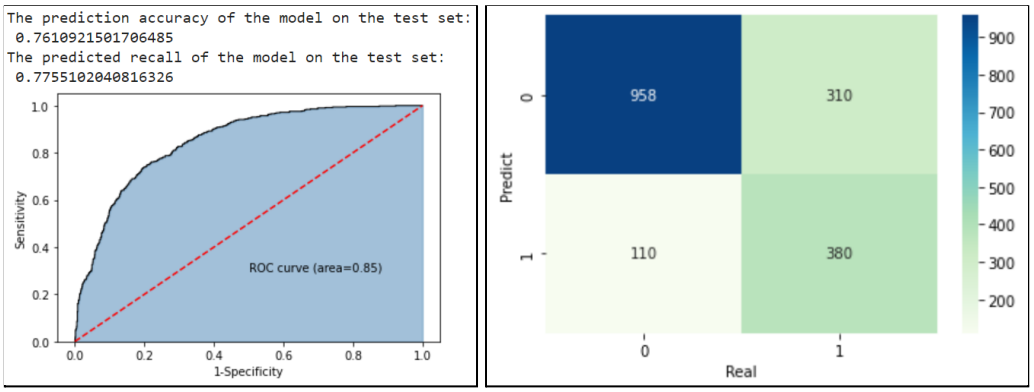

Figure 21: Accuracy, Recall, Confusion Matrix and AUC of Random Forest

11.3.2 Decision tree

Upon using grid search, it was found that Decision Trees's best parameter is {'max_depth': 8} with the best accuracy score of 0.83235.

The accuracy and recall that are derived from the confusion matrix are 0.72923 and 0.76327 respectively, while the AUC is 0.82.

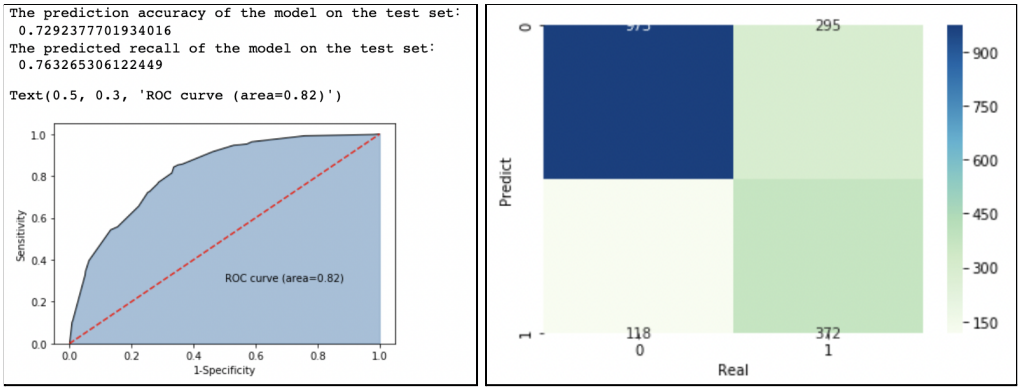

Figure 22: Accuracy, Recall, Confusion Matrix and AUC of Decision Tree

11.3.3 AdaBoost

Upon using grid search, it was found that AdaBoost's best parameter is {'n_estimators': 600} with the best accuracy score of 0.83902.

The accuracy and recall that are derived from the confusion matrix are 0.76962 and 0.78571 respectively, while the AUC is 0.85.

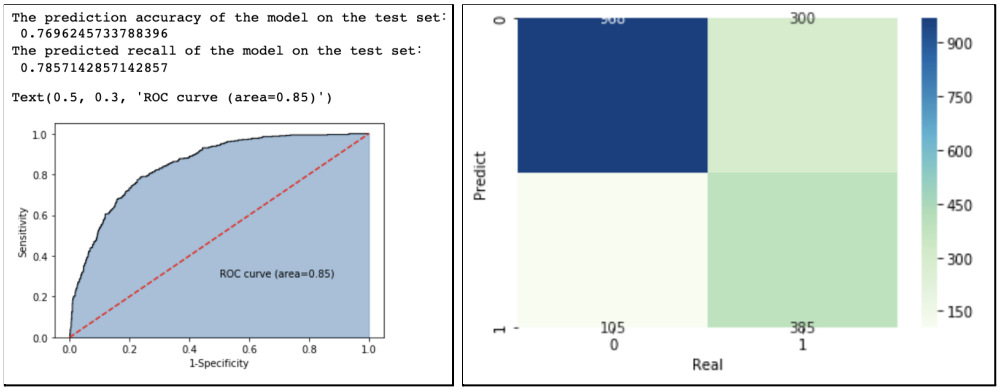

Figure 23: Accuracy, Recall, Confusion Matrix and AUC of AdaBoost

11.3.4 K-Nearest Neighbours (KNN)

Upon using grid search, it was found that KNN's best parameter is {'n_neighbors': 100} with the best accuracy score of 0.75892.

The accuracy and recall that are derived from the confusion matrix are 0.71388 and 0.85918 respectively, while the AUC is 0.85.

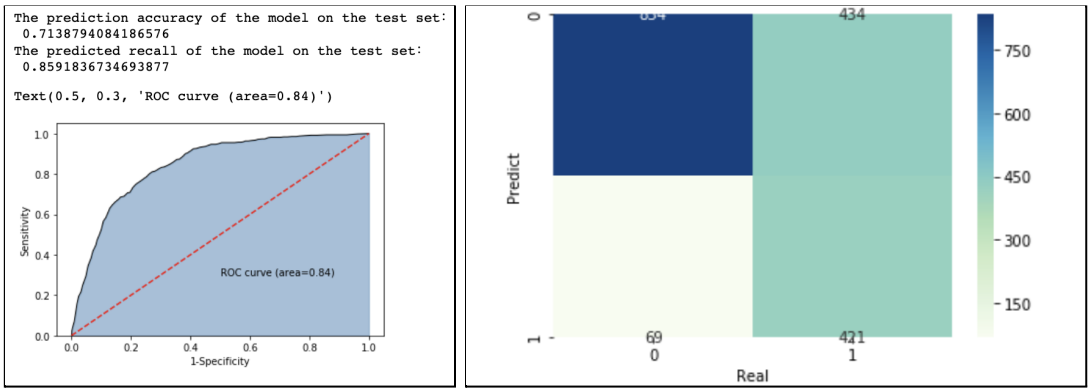

Figure 24: Accuracy, Recall, Confusion Matrix and AUC of KNN

11.3.5 Support Vector Machine (SVM)

For a linear hyperplane in SVM, upon using grid search, it was found that the best parameter is {'C': 0.8, 'gamma': 1.0 ,'kernel': 'linear'} with the best accuracy score of 0.81656.

The accuracy and recall that are derived from the confusion matrix are 0.75085 and 0.80816 respectively, while the AUC is 0.85.

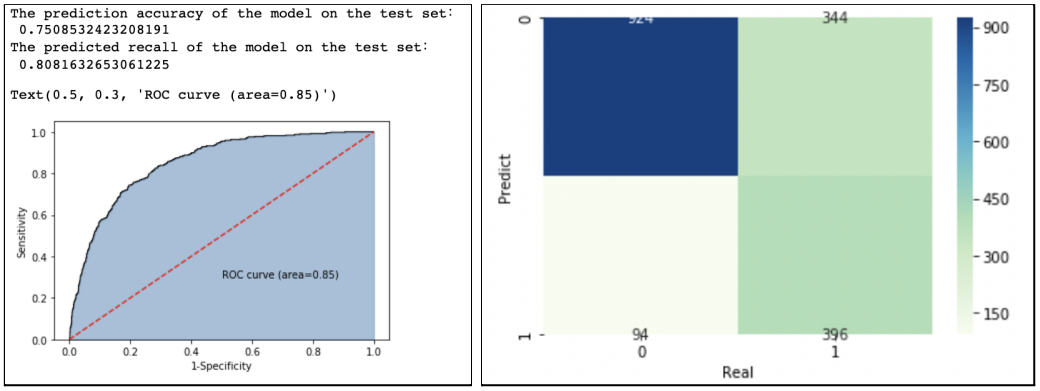

Figure 25: Accuracy, Recall, Confusion Matrix and AUC of SVM (Linear)

For non-linear hyperplane, radial basis function (rbf) in SVM, upon using grid search, it was found that the best parameter is {'C': 0.8, 'gamma': 1.0 ,'kernel': rbf} with the best accuracy score of 0.84968.

The accuracy and recall that are derived from the confusion matrix are 0.76222 and 0.51020 respectively, while the AUC is 0.78.

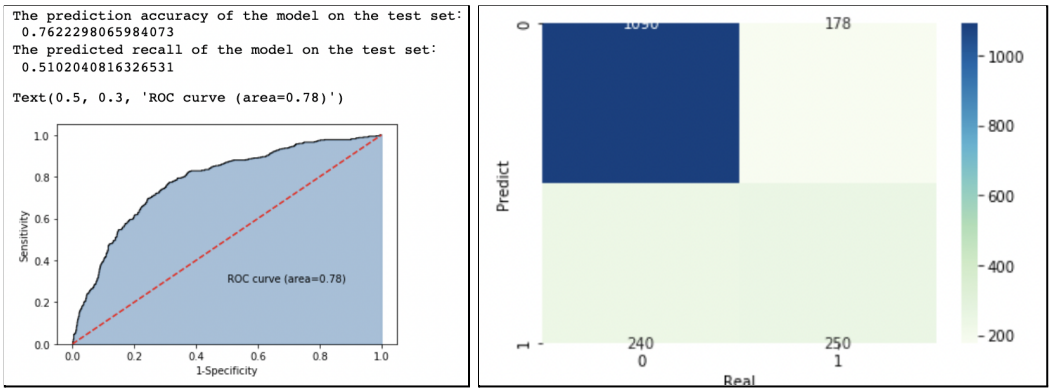

Figure 26: Accuracy, Recall, Confusion Matrix and AUC of SVM (Non-linear RBF)

11.3.6 LightGBM model

Upon using grid search, it was found that LightGBM's best parameter is {'n_estimators': 100} with the best accuracy score of 0.84994.

The accuracy and recall that are derived from the confusion matrix are 0.77019 and 0.72653 respectively, while the AUC is 0.85.

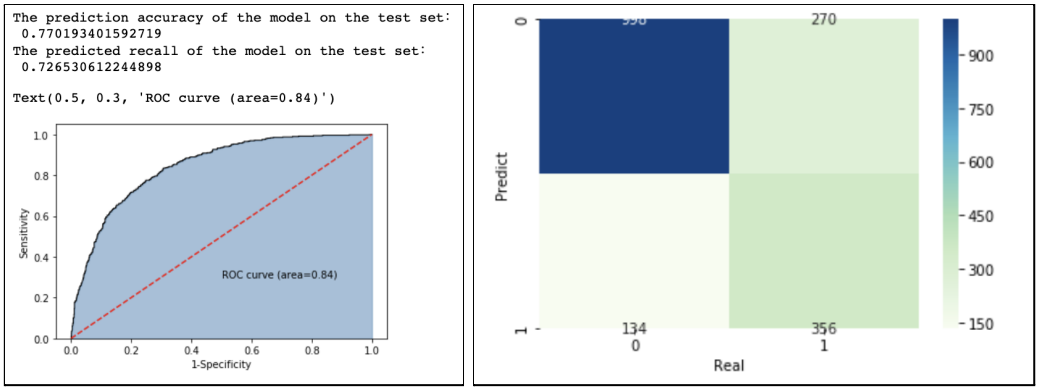

Figure 27: Accuracy, Recall, Confusion Matrix and AUC of LightGBM

11.3.7 XGBoost model

Upon using grid search, it was found that XGBoost's best parameter is {'n_estimators': 100} with the best accuracy score of 0.84994.

The accuracy and recall that are derived from the confusion matrix are 0.74118 and 0.78163 respectively, while the AUC is 0.85.

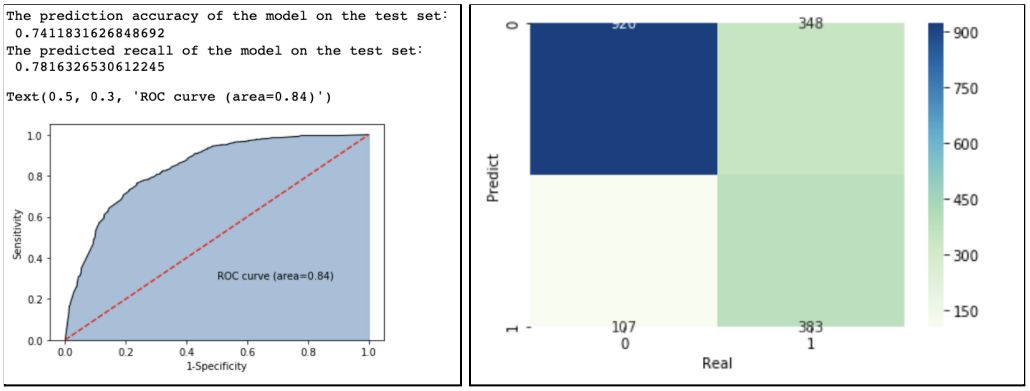

Figure 28: Accuracy, Recall, Confusion Matrix and AUC of XGBoost

11.3.8 Extra Tree

Upon using grid search, it was found that Extra Tree's best parameter is {'max_depth': 8, 'min_samples_leaf': 5, 'min_samples_split': 10, 'n_estimators': 600} with the best accuracy score of 0.79987.

The accuracy and recall that are derived from the confusion matrix are 0.76337 and 0.77551 respectively, while the AUC is 0.85.

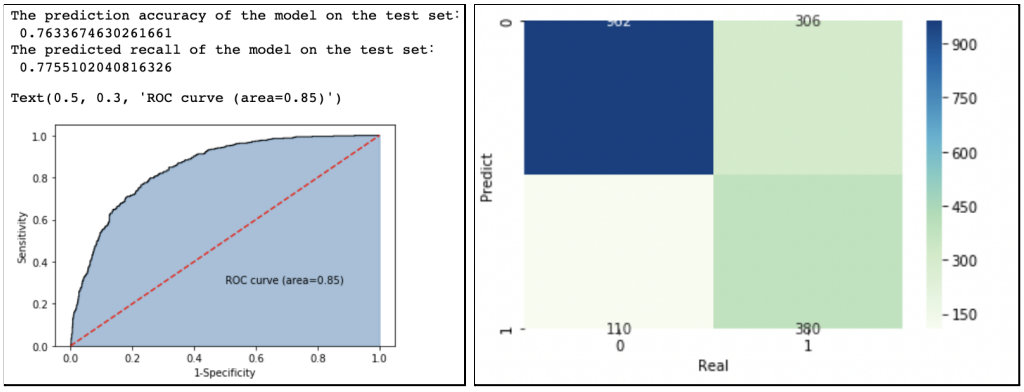

Figure 29: Accuracy, Recall, Confusion Matrix and AUC of Extra Tree

11.3.9 GBDT Gradient Boost Decision Tree

Upon using grid search, it was found that GBDT's best parameter is {'max_depth': 5, 'min_samples_leaf': 10, 'min_samples_split': 20, 'n_estimators': 100} with the best accuracy score of 0.77291.

The accuracy and recall that are derived from the confusion matrix are 0.75199 and 0.76122 respectively, while the AUC is 0.85.

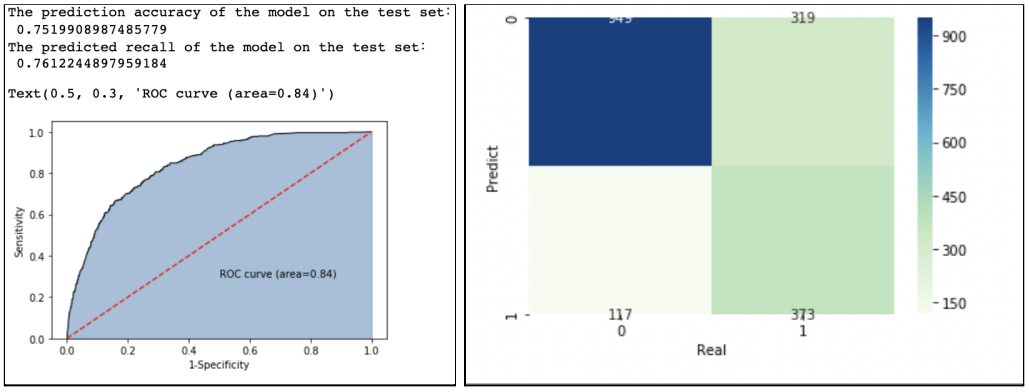

Figure 30: Accuracy, Recall, Confusion Matrix and AUC of GBDT Gradient Boost Decision Tree

11.3.10 Deep Learning ANN model

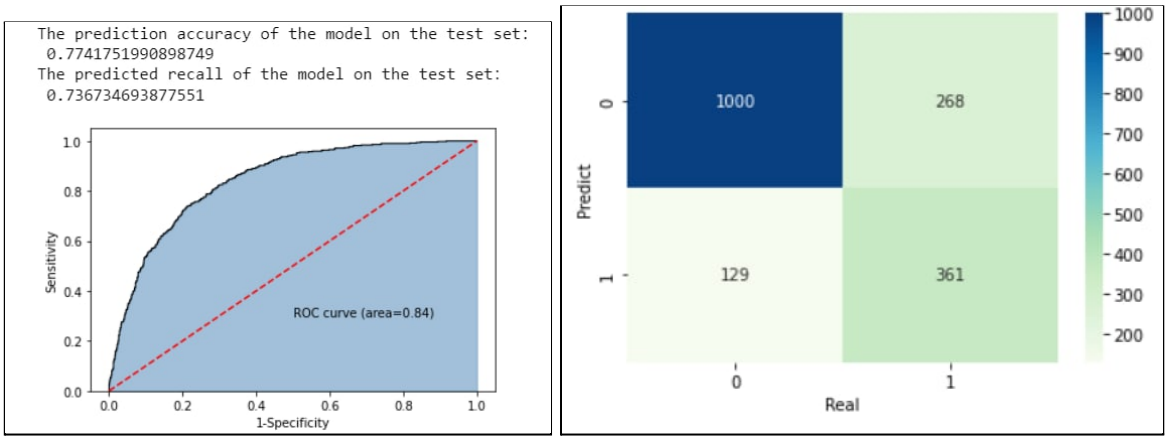

Figure 31: Accuracy, Recall, Confusion Matrix and AUC of Deep Learning ANN Model

11.4 Testing with Model fusion

11.4.1 Voting

The voting model integrates the predictions from Random Forest, Decision Tree, Adaboost, KNN, SVM (linear and non-linear), LightGBM, XGBoost, Extra Tree and GBDT and each model's contribution is weighted proportionally to its performance. Its best parameters are derived from the models it is made out of. The accuracy and recall that are derived from the confusion matrix are 0.72526 and 0.83878 respectively, while the AUC is 0.85.

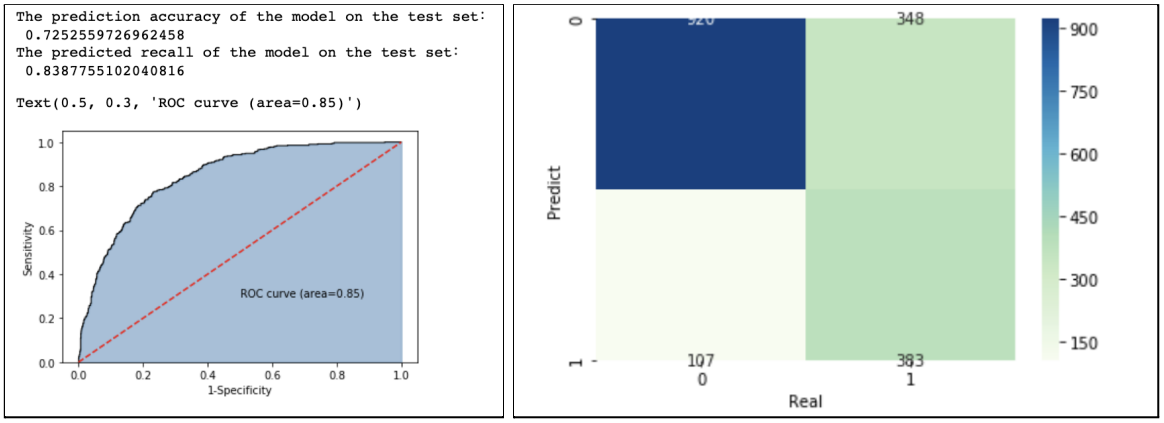

Figure 32: Accuracy, Recall, Confusion Matrix and AUC of Voting Model

11.5 Recommended Models and Performance Evaluation

Combined with the actual problem, the recommended model is AdaBoost, SVM with a linear hyperplane Voting model and random forest model as their AUC Area on the test set is the highest at 85% and out of the models with 75% AUC, the recommended models have the highest recall rates of 0.78571, 0.80816 and 0.83878 respectively.

These 3 algorithms have their respective reasons for performing better than the other 8 models. For Adaboost, the decision tree was used as a base to get a better and more stable result. This is better than single-class learning that learns a single target class only as Adaboost combines multiple weak classifiers (in this case, decision tree was used as a base) into a strong classifier for a better and more stable result.

For SVM Linear, it scales well with high dimensional data (large feature spaces), which suits our dataset with 20 features. It is also known to solve problems with small samples well, which suits out dataset with 7,000 records. It is also very suitable for binary classification, which in our case is "churn" or "non-churn". The strong generalisation ability is helpful in our case as the balance rate of any new dataset used in the future might be different. Lastly, it has no local minimum problem compared to algorithms such as neural networks.

For voting, it relies on the performance of many models, meaning that they will not be hindered by large errors or misclassifications from a single model. Poor performance from one model can be offset by a strong performance from other models which makes single learners unable to get a stable prediction result.

12. Conclusion

12.1 Conclusion

As mentioned, the ability to predict when or why a customer will churn will greatly impact the profit levels of telecom companies. Retaining customers is the key to a consistent and sustainable profit stream.

Through analysis, it is found that high churn users show the following characteristics: no partners or children, older, using fibre optic Internet service, additional entertainment services instead of guaranteed services, choosing monthly payment instead of annual payment, using electronic cheque methods, electronic bills and new users who have been with the company for less than half a year.

Through data mining, multiple classification prediction models were obtained, and the AUC value reached 0.85. Affected by the unbalanced data set, the accuracy of the model on the test set is not high, but the recall rate is above 80%, which sufficiently accounts for the vast majority of lost users.

12.2 Suggestions

It is important to utilise customer segmentation to better understand each customer group's purchasing patterns, as different types of users have different price sensitivities, to personalise each plan's payment amount and the benefits that should come with it. After understanding our customer persona, we have decided to address a few aspects of high churn customers.

Contracts and Payment Methods

In terms of contracts and payment methods, users should be encouraged to sign long-term contracts, preferably one or two-year packages, with added benefits or guaranteed value-added services to entice users to stay loyal. Telcos could also provide greater customization for their plans, allowing users to select and customise specific components of their plan. Surveys have shown that consumers are more inclined to companies that provide them with choice and control (Saratoga, 2017). By providing personalization options, customers are influenced to stay with the company, as all their needs and expectations can be easily met. More flexible payment options could also be made available to improve the payment experience of the customers.

User Types

In terms of user types, for older users who do not have a partner or children, they can launch preferential packages that are simple to understand and more affordable for seniors. These packages could also provide additional benefits such as express queues at the Telco shops or free subscriptions to streaming services to further attract seniors to their plans. Phones could also be bundled with these plans, which simplifies the process of setting up mobile phones for many seniors, thus making convenience one of the factors that retains older users. This will greatly encourage digital adoption among seniors, especially in a global world. For other users, tying guaranteed additional benefits like free streaming services with the plan will act as an enormous advantage as worldwide subscriptions to online video streaming has reached 1.1billion in 2020, and will only continue to increase. This could be an engagement driver, retaining the younger customers as well.

Network Services

In terms of network services, Telco companies should further investigate the quality of their optical fibre services, which branches out into two aspects: a. Service quality issues b. User satisfaction with the price of optical fibre services. High churn due to subscriptions to optical fibre services could be an indication of poor network quality or speed or unreasonable prices. When providing network services, Telcos should ensure that their networks are faster, reliable and more durable than their competitors, so they could provide the uncompromisable connectivity that many users need today. User surveys could also be conducted often to get a consensus from the Telco's consumers, and to gain insight on what exactly their consumers are unhappy with, so they could better improve their services.

Through the prediction model, Telco companies could better develop a plan to manage high churners of their company separately, and formulate better strategies for targeted packages and services that would likely convert these high churners into long-term users, positioning their company as being trustworthy to gain more loyal customers in the future.

12.3 Future Work

We have decided to address 2 aspects of our future work. The model aspect and also the problem aspect.

12.3.1 Model Aspect

Feature Engineering with WOC Encoding

We could apply feature engineering, namely feature binning with WOC encoding, to simplify and speed up data transformations, whilst enhancing the accuracy of our model. Feature engineering would increase suitability of our dataset to our models, thus providing better predictions.

Siamese Convolutional Neural Networks

We could also further enhance our models by using classification algorithms of imbalanced data based on differential siamese convolutional neural networks, instead of using GAN or SMOTE to deal with the data, like we did in this project. Siamese networks are networks that are made up of two or more identical networks that have identical weights. A pair of inputs are fed to these networks, and each network computes the features of one input. The similarity of the features are computed using either their difference or their dot product. Now, the classification problem has been converted to a similarity problem. Siamese networks focus on learning embeddings (deeper layers) that place similar classes and concepts together. The network is trained to minimise the distance between samples of the same class and increases the inter-class distance.

Siamese networks are more robust to class imbalance, just giving a few images per class is sufficient for the Siamese networks to learn and recognize them in the future. It can also perform better than taking the average of 2 correlated supervised models. Currently, our dataset is not very imbalanced, which is why GAN and SMOTE are not as effective in this dataset, thus we could have ventured further using Siamese Neural Networks.

12.3.2 Problem Aspect

Social Data on Customers

We could gather more social data on our customers, as customers' social relationships influence the decision making process of choosing a Telco company or plan. There is a high propensity of a customer to churn if their social network has also recently churned, which shows how a customer's purchasing decision could be dependent on their family and friends. We could utilise a novel method to extract the dynamic relevance of each customer using social network analysis techniques, and understand more about the underlying relationships between social ties and churning from Telco companies. This could help us to build a stronger predictive model.

Unsupervised Learning (Clustering)

We could also explore unsupervised learning, namely clustering. Clustering gives us a very good understanding and summarization of data, and helps us understand any meaningful intuition of our data's underlying structure. Clustering has the ability to group similar data points together, so that we can group our customers by similar characteristics, and further develop different models for these different clusters, targeting the different customer segments deeper and accurately.

13. References

Calzada-Infante, Laura & Óskarsdóttir, María & Baesens, Bart. (2020). Evaluation of customer behavior with temporal centrality metrics for churn prediction of prepaid contracts. Expert Systems with Applications. 160. 113553. 10.1016/j.eswa.2020.113553.

Campbell, P. (2021). Customer Churn Models: Lowering CAC, Maximizing Retention. ProfitWell. Retrieved January 22, 2022, from https://www.profitwell.com/customer-churn/models

Dorard, L., & Patel, N. (2021). *How to Improve Your Subscription Based Business by Predicting Churn*. Neil Patel. Retrieved January 22, 2022, from https://neilpatel.com/blog/improve-by-predicting-churn/

Dutt, A. (2021, March 11). Siamese Networks Introduction and Implementation - Towards Data Science. Medium; Towards Data Science. https://towardsdatascience.com/siamese-networks-introduction-and-implementation-2140e3443dee

KDnuggets. (2017, March 1). *What is Customer Churn Modeling? Why is it valuable?* KDnuggets. Retrieved January 22, 2022, from https://www.kdnuggets.com/2017/03/datascience-customer-churn-modeling.html

Li, Bo & Xie, Jiuzuo. (2020). Study on the Prediction of Imbalanced Bank Customer Churn Based on Generative Adversarial Network. Journal of Physics: Conference Series. 1624. 032054. 10.1088/1742-6596/1624/3/032054.

N. Lu, H. Lin, J. Lu and G. Zhang, "A Customer Churn Prediction Model in Telecom Industry Using Boosting," in *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1659-1665, May 2014, doi: 10.1109/TII.2012.2224355.

New Consumer Survey Shows U.S. Mobile Carriers Are Missing The Mark With Unlimited Offerings | MATRIXX Software. (2018). MATRIXX Software. https://www.matrixx.com/press_release/new-consumer-survey-shows-u-s-mobile-carriers-missing-mark-unlimited-offerings/

Saratoga, C. (2017, May 24). *New Consumer Survey Shows US Mobile Carriers Are Missing The Mark With Unlimited Offerings*. MATRIXX Software.

https://www.matrixx.com/press_release/new-consumer-survey-shows-u-s-mobile-carriers-missing-mark-unlimited-offerings/

W. Bi, M. Cai, M. Liu and G. Li, "A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn," in IEEE Transactions on Industrial Informatics, vol. 12, no. 3, pp. 1270-1281, June 2016, doi: 10.1109/TII.2016.2547584.