# 1. Introduction

## 1.1 Problem Overview

## 1.1.1 Domain Background

This project was inspired by the Kaggle competition [Customer Churn Prediction 2020](https://www.kaggle.com/competitions/customer-churn-prediction-2020/overview). This competition is about predicting whether a customer will change telecommunications provider, something known as "churning". The reason why this is important is that the technical progress and the increasing number of operators globally have made the industry competitive [1]. Companies are working hard to survive in this competitive market depending on multiple strategies.

There are often three main strategies for generating more revenue within a business [2]:

1. Acquiring new customers
2. Upselling existing customers
3. Increase the retention period of customers

However, comparing these strategies taking the value of return on investment (RoI) of each into account has shown that the third strategy is the most profitable strategy. The reason being is that retaining an existing customer costs much lower than acquiring a new one, in addition to being considered much easier than the upselling strategy. To apply the third strategy, we need to decrease the potential of customer churn by putting systems in place to do so. Hence why exploring machine learning techniques for predicting customer churn can provide huge financial benefits to companies.

### 1.1.2 Related Work

Many approaches were applied to predict churn in telecom companies. Most of these approaches have used machine learning and data mining. The majority of related work focused on applying only one method of data mining to extract knowledge, and the others focused on comparing several strategies to predict churn.

Gavril et al. [3] presented an advanced methodology of data mining to predict churn for prepaid customers using a dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. The author used AUC to measure the performance of the algorithms. The AUC values were 99.10%, 99.55% and 99.70% for Bayes Networks, Neural networks and support vector machine, respectively.

He et al. [4] proposed a model for prediction based on the Neural Network algorithm in order to solve the problem of customer churn in a large Chinese telecom company which contains about 5.23 million customers. The prediction accuracy standard was the overall accuracy rate, and reached 91.1%.

Idris [5] proposed an approach based on genetic programming with AdaBoost to model the churn problem in telecommunications. The model was tested on two standard data sets. One by Orange Telecom and the other by cell2cell, with 89% accuracy for the cell2cell dataset and 63% for the other one.

Huang et al. [6] studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhances the process of predicting churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China’s largest telecommunications company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC.

Ahmad et al. [7] investigated the churn problem by developing a model using machine learning techniques on a big data platform and building a new way of feature engineering and selection. Their study measured the performance of the model using the AUC and obtained an AUC value of 93.3% after applying the XGBOOST algorithm.

Lalwani et al. [8] tacked the churn prediction problem by using a gravitational search algorithm for feature selection while several machine learning models were applied, namely, logistic regression, naive bayes, support vector machine, random forest, decision trees. The highest AUC score of 84% was achieved by both Adaboost and XGboost classifiers.

### 1.1.3 Datasets and Inputs

The training dataset contains 4250 samples. Each sample contains 19 features and 1 boolean variable churn which indicates the class of the sample. The 19 input features and 1 target variable are:

**File Descriptions**

All of these files can be found in dataset folder within the submission:

* **train.csv** - the training set. Contains 4250 lines with 20 columns. 3652 samples (85.93%) belong to class churn=no and 598 samples (14.07%) belong to class churn=*yes*
* **test.csv** - the test set. Contains 750 lines with 20 columns: the index of each sample and the 19 features (missing the target variable churn).
* **sampleSubmission.csv** - a sample submission file in the correct format

**Data Fields**

1. state, *string*. 2-letter code of the US state of customer residence
2. account\_length, *numerical*. Number of months the customer has been with the current telco provider
3. area\_code, *string*=area\_code\_AAA where AAA = 3 digit area code.
4. international\_plan, *string*,(yes/no). The customer has international plan.
5. voice\_mail\_plan, *string*, (yes/no). The customer has voice mail plan.
6. number\_vmail\_messages, *numerical*. Number of voice-mail messages.
7. total\_day\_minutes, *numerical*. Total minutes of day calls.
8. total\_day\_calls, *numerical*. Total minutes of day calls.
9. total\_day\_charge, *numerical*. Total charge of day calls.
10. total\_eve\_minutes, *numerical*. Total minutes of evening calls.
11. total\_eve\_calls, *numerical*. Total number of evening calls.
12. total\_eve\_charge, *numerical*•. Total charge of evening calls.
13. total\_night\_minutes, *numerical*. Total minutes of night calls.
14. total\_night\_calls, *numerical*. Total number of night calls.
15. total\_night\_charge, *numerical*. Total charge of night calls.
16. total\_intl\_minutes, *numerical*. Total minutes of international calls.
17. total\_intl\_calls, *numerical*. Total number of international calls.
18. total\_intl\_charge, *numerical*. Total charge of international calls
19. number\_customer\_service\_calls, *numerical*. Number of calls to customer service
20. churn, *categorical*, (yes/no). Customer churn - target variable.

## 1.2 Problem statement

### 1.2.1 Problem

Customer churn is a major problem and one of the most important concerns for large companies. Due to the impact customer churn has on the revenues of companies it is important to develop means to predict customer churn.

Customer churn is a common problem for businesses that provide subscription services. It is often difficult to determine when exactly a customer is likely to churn, due to the fact that it could be caused by several reasons.

The challenge is to predict whether a customer will churn or not using 19 input features defined in the dataset. The area under the curve (AUC) score of the model will be the main metric for determining its success.

### 1.2.2 Solution

With the advancements in artificial intelligence, the possibility to predict customer churn has increased significantly.

Having looked at the current literature, most research has been focused on using single estimators for making predictions. This project intends on taking an alternative approach by building an ensemble voting classifier comprised of 3 models that have the highest AUC. I will use AWS AutoGluon to find the top 3 models. The models that I will use are:

1. **LightGBM:** an implementation of the gradient-boosted trees algorithm that adds two novel techniques for improved efficiency and scalability: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).
2. **CatBoost:** an implementation of the gradient-boosted trees algorithm that introduces ordered boosting and an innovative algorithm for processing categorical features.
3. **XGBoost:** an implementation of the gradient-boosted trees algorithm that combines an ensemble of estimates from a set of simpler and weaker models.
4. **Random Forests:** Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.
5. **K Nearest Neighbours:** a non-parametric method that uses the k nearest labelled points to assign a label to a new data point for classification or a predicted target value from the average
6. **Linear Learner Algorithm** learns a linear function for regression or a linear threshold function for classification.

Once the best model with the highest AUC is determined the model will then be registered and deployed to an AWS SageMaker endpoint.

### 1.2.3 Solution workflow:

The workflow for approaching a solution:

1. **Data Analysis:** understand the datasets
2. **Features Transformation:** convert variables into features. Standardize/normalize features, apply numerical transformations
3. **Features Selection:** select relevant features
4. **Machine Learning Models:** Train different models and perform hyperparameter tuning
5. **Evaluation:** evaluate the performance of each model, and check the possibilities of combining them to achieve an optimal model
6. **Deployment:** deploy the trained model to an AWS endpoint

## 1.3 Evaluation metrics

After a quick look at the dataset, I noticed that the dataset was imbalanced. Moreover, for business reasons, the model must be able to predict each class very well, therefore the Area Under the Curve (AUC) score will be used for evaluating the model. Additionally, the decision tree diagram below further justifies why AUC score would be the most relevant metric to use to evaluate the model because we will need probabilities to determine the propensity to churn and both classes are equally important. AUC score has also been used for churn prediction analysis in other studies [8][7].

Diagram

Description automatically generated

Figure 1: Decision tree diagram for choosing a binary classification performance metric. Image source [9]

# 2. Method

## 2.1 Data Pre-processing

### 2.1.1 Loading and cleaning the dataset

As I loaded the dataset, I checked for any null values and inconsistent data types. Upon looking at the dataset, one can see that each column has an equal amount of non-null values, indicating that there are no instances of missing data. However, the data types in the dataset are of type object, int64, and float64 indicating varying data types and a mixture of categorical and numerical variables. Lastly, I checked for duplicate values and did not find any.

### 2.1.2 Creating new input features

Upon looking at the dataset, I created additional features total\_minutes, total\_calls and total charges to observe their correlations with churn and their impact on the model's performance.

Next, I one-hot encoded the categorical features to make all the features numerical so that I could create a correlation matrix of all features. Lastly, I label encoded the target variables where 0 = ‘not churned’ and 1 = ‘churned’.

### 2.1.3 Train/Validation Split

### I created a validation set in order to evaluate the performance of the model and to see how well the model could generalise on unseen data. The training and validation sets were split within an 80/20 split in a stratified fashion.

## 2.2 Analysing and visualising the dataset

### 2.2.1 Visualising the class distribution

I first visualised the class distribution to have a better understanding of the split between churned and not churned. From figure 2, we can see that 85.94% of the dataset was classified as not churned, indicating that the dataset was highly imbalanced.

### Chart Description automatically generated

Figure 2: Class Frequency distribution of the Churn Dataset

### 2.2.2 Visualising the distributions of the numerical features

### Most of the features tend to form a normal distribution. However, number\_service\_calls, total international calls, number\_vmail\_messages are more right skewed than others.

### Chart, box and whisker chart Description automatically generated

Figure : Histograms of the numerical features within the dataset

### 2.2.3 Correlation Matrix

### To have a better understanding of the relationship between the features and the target variable I created two correlation matrices. One for the top 10 features with a positive correlation with churn and another for the top 10 features with a negative correlation with churn. As we can see from figure 4 below, “international\_plan\_yes” is the feature with the highest correlation with churn followed by “total\_charge”. This is not surprising as it could be expected that increase in charges could cause a customer to switch to another provider that could offer a better deal. Moreover, total\_day\_minutes, total\_day\_charge and total\_minutes all have a high correlation with one another, thus suggesting multicollinearity. This could be problematic for a logistic regression model however I’ll also be training other tree-based models so they shouldn’t be too affected by this. On the other hand, I noticed that “voice\_mail\_plan\_yes” had the lowest correlation with churn, followed by number of vmail messages. Not sure why that would be the case but interesting to observe.

### Chart Description automatically generated

Figure 4: Top 10 features with a positive correlation with churn

### Chart Description automatically generated

Figure 5: Top 10 features that have a negative correlation with churn

### 2.2.4 Pair Plot

### Visualising the pair plot was very useful for having a good understanding of the pairwise relationship between the numerical features. We can see that evening\_minutes and evening\_charge features have a very strong positive correlation. Moreover, as seen earlier, all the numerical features tend to form a normal distribution

### Shape, polygon Description automatically generated

Figure : Pairplot of all the numerical features vs churn. The diagonal plots are a univariate distribution plot drawn to show the marginal distribution of the data in each column

## 2.3 Feature selection

In order to pick the best features, I created 3 different datasets and then trained a model using 5-fold cross-validation to find the best dataset to use for hyperparameter tuning.

The datasets that I created were:

* 1. Full dataset – all features
  2. Top 10 features that correlated with churn
  3. Recursive Feature Elimination (RFE)

Additionally, RFE is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model’s **coef\_** or f**eature\_importances\_** attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.

In order to optimise RFE, I chose different base models to implement the feature selection and then trained and computed the AUC scores for the different datasets using the LightGBM model. I found that the XGBoost model had the highest AUC score on average, with a score of 94%. (see Figure 7).

Chart, box and whisker chart

Description automatically generated

Figure : Boxplots of the AUC scores of the 4 base models that were used for RFE optimisation

Moreover, to finally select the features, I used XGBoost as the base model and found that the seven features it selected were:

* + 1. Number\_vmail\_messages
    2. Total\_intl\_minutes
    3. Total\_intl\_calls
    4. Number\_customer\_service\_calls
    5. Total\_charge
    6. Area\_code\_area\_code\_415
    7. International\_plan\_no

## 2. 4 Train the models

### 2.4.1 Models

### The models that I used in the project were:

### LightGBM: an implementation of the gradient-boosted trees algorithm that adds two novel techniques for improved efficiency and scalability: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

### Catboost an implementation of the gradient-boosted trees algorithm that introduces ordered boosting and an innovative algorithm for processing categorical features.

### 2.4.2 Find the top-performing dataset

### In order to find the top-performing dataset, I trained a default LightGBM model and evaluated the AUC scores with 5-fold cross-validation (see Figure 8) and found that the RFE dataset had the highest average AUC score of 93.8%.

### Chart, box and whisker chart Description automatically generated

Figure : Box plot showing LightGBM's training performance across the 3 datasets with 5 fold-cross validation

### 2.4.2 Find the top-performing model

### Chart, box and whisker chart Description automatically generated

Figure : Box plots of the AUC scores of six machine learning models trained on the RFE datasets with 5-fold cross-validation

### Next, I trained and scored six models with their default hyperparameters using the RFE dataset to find the top performing model.

### 2.4.3 Hyperparameter tuning of top models

# 3. Evaluation on Validation set

3.1 AUC scores

3.2 Confusion Matrices

3.3 ROC Curve

Chart, line chart

Description automatically generated

# 4. Kaggle Submission

# 5. Model deployment in Sagemaker