**1. Intro**

Before diving into this project, it is important to know what customer churn is. Customer churn refers to the rate at which customers will stop doing business with a company in a given period of time, where the “doing business” part differs from industry to industry. For this project, the focus will be on the telecommunication industry, where customer churn typically represents the percentage of customers (also referred to as subscribers) who either cancel their service contracts (phone plan, internet service etc.) or switch to a competitor.

After a quick Google search, I managed to find a list of telecommunication companies in Denmark, which shows a list of 34 (!) different companies[[1]](#footnote-1). Such a high number of companies offering the same or similar services makes for a highly competitive industry with companies always fighting to have the cheapest prices, best service, or a combination of both, to attract new customers. In such a highly competitive industry customer churn is a critical metric for a few reasons such as the following.

1. **Revenue loss:**

There is a direct link to be made between customer churn and loss of revenue. In this case the revenue is subscription revenue, which is the type of revenue the telecommunication industry relies on, unlike other industries where revenue might come from one time sales. If a telecommunication company loses a significant number of customers, it can create a rather large gap that will hurt expected revenue and will take time to fill.

1. **Customer acquisition cost:**

Acquiring a new customer can be rather expensive since it involves marketing, advertising, promotions, possibly also discounts and more to attract new customers from competitors. According to a Harvard Business Review article[[2]](#footnote-2), acquiring new customers can be 5 to 25 times more expensive than retaining customers, depending on what industry you are in. The article also suggests that increasing customer retention by as much as 5% can increase profits by 25% to 95%. So, minimizing customer churn can prove to be valuable for telecommunication companies.

1. **Loss of market share:**

It can be argued that the telecommunication market is saturated, since it to a certain extent is all about customers changing supplier, and not really about getting new customers into the market. Since the number of new potential customers is somewhat limited, a high churn rate can lead to a loss of market share, since the company would suddenly sit on a smaller number of customers in the saturated market, leading to higher market shares for their competitors.

1. **Brand Perception:**

If a company, in any market really, is experiencing a consistent customer churn, it can signal issues customers might what to avoid, such as poor service, inadequate customer service/support, or prices that are higher than similar product on the market. Such issues can lead to dissatisfaction among customers, which in turn can lead to bad word-of-mouth and a damaged reputation, which can make it harder to attract new customers. On the other hand, retaining customers and offering good service, good customer support and some competitive prices can foster loyalty, which in turn can lead to good reputation and good word-of-mouth. Good reputation can also make it easier to both retain customers and attract new customers.

The aim of this project is to come up with a tool that, to some extent, can predict customer churn. Such a tool can be a powerful tool for addressing and understanding customer churn, and in turn hopefully help develop retention strategies.

By applying this model/tool to large amounts of customer data, it can help identify possible key features/drivers that affect customer churn, such as frequent complaints, late payments, or issues related to usage and service. With this kind of data insight telecommunication companies can focus resources on the core of the issues leading to higher churn rates.

By utilizing a machine learning model, the tool can be customized to fit specific needs. For example, a large company with a large number of customers might want to divide their customers into segments based on churn risk, such as low-risk/loyalty, medium-risk and high-risk. By dividing customers into segments, the company can develop a specific retention strategy for each segment, since it is very likely that each segment will have different wants and needs to stay with the company. For example, cost-sensitive customers could be more interested in bundle offers or discounts, while unhappy customers could be more interested in upgrades in service or support.

To further develop this tool real-time data processing can be introduced, so the company can predict churn before it happens, and thus try to predict what to offer and when to offer it. An example of this continuous behavior monitoring could be if a customer suddenly uses significant less data or there is a significant increase in customer support tickets.

So, all in all, machine learning powered churn prediction can add several benefits to telecommunication companies operating in a saturated market, all for improving profitability and competitive advantage.

**2. Theory and methodology**

Customer churn analysis leverages machine learning models to identify patterns in order to predict customer behavior. An approach like customer churn analysis enables a business, in this case a telecommunication company, to take proactive action in order to reduce customer churn.

For this project the theoretical foundation supervised machine learning models, different feature engineering and data preparation techniques, and some understanding of customer behavior. The aim of this chapter is to elaborate on the theoretical foundation of the project, trying to explain certain defining steps.

* **Supervised Machine Learning**

Traditionally models such as Random Forest and logistic regression are very popular options for predicting customer churn. These two models each have their own benefits over the other, such as Random Forest being very effective when working with large datasets, especially if the datasets contain numerous features. One very big benefit of Random Forest is its ability to reduce overfitting despite working with the aforementioned large datasets. Logistic regression on the other hand offers simplicity and excellent interpretability making it easy it identify key drivers for customer churn.

For this project I have explored the use of the gradient boosting algorithm, CatBoost.

The data used in this project would be considered categorical, making CatBoost a good choice for the predictive part, since it is optimized for working with categorical data while not needing extensive preprocessing. On top of that, its ability to naturally integrate both numerical and categorical features, combined with high efficiency, further proves CatBoost to be a solid choice for customer churn prediction.

* **Feature engineering and data preparation**

When building accurate and reliable machine learning models it is important to do effective feature engineering and data preprocessing. This section will go through the steps taken in this project.

* *Handling missing values*

In the dataset there is a column called “TotalCharges” that contains missing values. To get around this the data in the column was converted to numerical values and the missing values was imputed as a product of the columns “MonthlyCharges” and “tenure”. This ensures a complete and meaningful dataset for the machine learning model.

* *Standardizing categorical data*

Multiple columns contain inconsistent categorical data, which makes it harder to train a model in a meaningful and effective way. Because of that, categorical values were converted to “Yes” and “No” to simplify the data and improve the model’s ability to interpret the data. For instance, “No phone service” was changed to “No” in the column “MultipleLines”. These columns also had categorical data converted “OnlineSecurity“, “OnlineBackup“, “DeviceProtection“, “TechSupport“, “StreamingTV“, “StreamingMovies“.

* *Encoding target variable*

Churn, as a target variable, was converted from a categorical Yes/no to a numeric 1/0, to ensure that the model had the right data type for the classification tasks at hand.

* *Stratified sampling for train-test splitting*

Since churned customers are fewer the non-churned customers in the data set, it is considered imbalanced and thus needs some work before it can be split for the training and testing of the model. To do this, a StratifiedShuffleSplit was done to ensure a more proportionally correct representation of the “Churn” data in both the train and test split. This helps by improving the model’s performance in generalizing when applied to unseen data.

* *Identifying categorical features for modelling*

Since this project is using the CatBoostClassifer model, which natively handles categorical data, there is no need for traditional encoding such as One-Hot encoding. What it needs is for the categorical columns in the dataset to be explicitly defined. This approach the preprocessing steps while not negatively impacting computational efficiency.

* **Deployment**

For deployment of the model a few neat tools have been used, such as Streamlit and FastAPI. Streamlit makes it easy to quickly develop user-friendly interactive dashboards/apps, so naturally that was an easy choice for the front-end part of the project. The app enables the telecommunication company to input customer data and get real-time customer churn analysis.

To keep the future possibilities a bit more open ended, it was decided to expose the model using FastAPI, opening up for the possibility to integrate the model in automated churn prediction workflows on other systems/platforms.

1. https://mobil-daekning.dk/oversigt-danske-mobilselskaber/ [↑](#footnote-ref-1)
2. https://hbr.org/2014/10/the-value-of-keeping-the-right-customers [↑](#footnote-ref-2)