Asian Institute of Technology

TELECOMMUNICATION CUSTOMER CHURN PREDICTION WITH MACHINE LEARNING

by

Kaung Nyo Lwin

Patsachon Pattakulpong

A Report Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Data Science and Artificial Intelligence

CHAPTER 1

INTRODUCTION

In the highly competitive telecommunications industry, customer retention has emerged as a critical factor for sustaining business growth and ensuring profitability. The phenomenon of customer churn, defined as the rate at which users discontinue their subscriptions, represents a significant challenge to revenue stability. Retaining existing customers is generally more cost-effective than acquiring new ones, emphasizing the financial implications of high churn rates.

Furthermore, elevated churn rates often indicate deeper organizational or service-related issues, such as subpar service quality, declining customer satisfaction, or intensified competition in the market. These challenges underscore the importance for telecom providers to adopt proactive strategies for churn management. By addressing the root causes of churn, companies can safeguard their customer base, maintain profitability, and enhance their market position.

1.1 Business Perspective on Customer Churn in the Telecommunication Industry

From a business standpoint, reducing customer churn is essential for sustaining profitability and ensuring long-term success. Retaining existing customers is notably more cost-effective compared to acquiring new ones, making churn reduction a critical focus area for telecom companies. This strategy directly influences Customer Lifetime Value (CLTV), a vital metric for measuring the long-term financial contribution of customers. A higher CLTV translates to greater revenue generation over time, as loyal customers tend to make repeated purchases and engage more with the company’s services.

High churn rates often reflect underlying issues such as poor customer satisfaction, which can erode brand loyalty and negatively impact a company's reputation. Addressing churn proactively allows businesses to build a stronger brand image, enhance customer trust, and promote loyalty. Moreover, satisfied customers are more likely to engage in positive word-of-mouth, which can attract new customers and reinforce market presence. In today’s competitive telecommunication landscape, managing churn effectively is indispensable for maintaining a strong market position. Customers are increasingly drawn to competitors offering better pricing, improved service quality, or enhanced customer experiences. Therefore, reducing churn not only preserves the existing customer base but also prevents potential revenue losses while reinforcing the company’s competitive edge in the industry.

1.2 Importance of Telecommunications Customer Churn Projects

Addressing customer churn in the telecommunications industry is critical for achieving business sustainability and growth in a highly competitive market. Churn management projects offer significant strategic benefits, including the following

1.2.1 Proactive Customer Retention

Churn prediction models empower telecom companies to identify at-risk customers early in their journey. With this knowledge, businesses can implement targeted retention strategies, such as personalized incentives, service plan enhancements, or tailored customer support. This proactive approach minimizes customer loss and strengthens customer loyalty.

1.2.2 Cost Efficiency

Retaining existing customers is far more economical than acquiring new ones. Customer acquisition involves significant costs in marketing, onboarding, and promotions, whereas reducing churn allows businesses to allocate resources more efficiently. By focusing on customer retention, telecom companies can lower their operational costs and improve profitability.

1.2.3 Increased Customer Lifetime Value (CLTV)

Churn reduction efforts directly enhance Customer Lifetime Value (CLTV), a critical metric for long-term business success. Retained customers tend to contribute higher revenue over time through continued use of services, upgrades, and additional purchases. A higher CLTV translates into sustained business growth and financial stability.

1.2.4 Enhanced Customer Experience

Understanding the root causes of customer churn provides telecom companies with actionable insights into customer behavior and preferences. By addressing these issues, companies can improve service quality, optimize product offerings, and meet customer expectations more effectively. These improvements lead to higher satisfaction and loyalty, further reducing churn rates.

CHAPTER 2

PROBLEM STATEMENT

The telecommunications industry faces significant challenges due to increasing customer churn, which has a direct and adverse impact on revenue and profitability. In an environment where customer loyalty is declining and switching costs are minimal, telecom companies are experiencing rising churn rates. This issue not only leads to substantial financial losses but also increases the costs associated with acquiring new customers. While companies invest heavily in marketing and customer service initiatives, the inability to accurately predict and understand the underlying causes of churn limits their ability to implement effective and proactive retention strategies.

This project addresses the pressing need for a solution by proposing the development of a robust predictive model. The model aims to accurately identify customers at risk of churning by analyzing a range of critical factors, including usage patterns, customer demographics, service quality, and historical churn data. By utilizing advanced machine learning techniques, the model seeks to enhance the company's capacity to:

2.1 Implement targeted retention strategies tailored to at-risk customers.

2.2 Optimize marketing resources, reducing unnecessary expenditures.

2.3 Lower churn rates while simultaneously improving customer satisfaction.

Through these efforts, the project ultimately aspires to contribute to long-term revenue growth and increased business sustainability. By bridging the gap between understanding churn dynamics and actionable intervention, the proposed solution provides a strategic pathway for addressing one of the most pressing challenges in the telecommunications sector.

CHAPTER 3

RELATED WORK

Predicting customer churn in the telecommunications industry has been a widely researched area, with numerous approaches utilizing machine learning and data mining techniques. These efforts have predominantly focused either on applying a single method for extracting knowledge or on comparing multiple strategies to identify the most effective approach for churn prediction.

3.1 Logistic Regression

Logistic Regression is one of the most commonly used techniques in churn prediction due to its simplicity and interpretability. It is particularly effective when a baseline accuracy is needed. Studies, such as those by **Tsai and Lu (2009)**, employed Logistic Regression to model customer churn, emphasizing feature importance to understand the key drivers of churn. Factors such as contract duration, data usage, and service complaints were identified as critical churn indicators. However, Logistic Regression is limited in its ability to capture complex, non-linear relationships in customer behavior, which can reduce its effectiveness in more intricate scenarios.

3.2 Decision Trees and Random Forests

Decision Trees and Random Forests are popular techniques in telecom churn prediction due to their ability to model non-linear relationships and provide interpretable results. **Idris et al. (2012)** demonstrated the **effectiveness of Random Forests in predicting customer churn**, achieving higher accuracy than Logistic Regression models. By constructing multiple decision trees, Random Forests help mitigate the overfitting problem commonly associated with single decision trees. Furthermore, these methods excel in handling large, unbalanced datasets, making them particularly suitable for telecom churn prediction tasks.

3.3 Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) have gained recognition for their exceptional performance in customer churn prediction tasks. **Chen and Guestrin (2016)** utilized **XGBoost**, a highly optimized GBM implementation, to predict churn. The iterative improvement of weaker models in GBM results in high predictive accuracy. These methods have consistently outperformed other techniques in churn prediction competitions and real-world applications. However, they require significant computational power and careful parameter tuning to achieve optimal results.

3.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) have also been employed for telecom churn prediction, as seen in the work of **Coussement and Van den Poel (2008).** Their study showed that SVM outperformed traditional methods such as Logistic Regression and Decision Trees by modeling complex, non-linear decision boundaries between churn and non-churn customers using kernel functions. Despite its effectiveness, SVM is computationally intensive for large datasets and requires careful hyperparameter tuning, such as regularization parameters and kernel selection.

3.5 Addressing Dataset Imbalance

Customer churn datasets are often imbalanced, with fewer churned customers compared to non-churned ones. This imbalance can adversely affect the performance of machine learning models. To address this, **Chawla et al. (2002)** introduced the **Synthetic Minority Over-sampling Technique (SMOTE)**, which generates synthetic samples of the minority class to balance the dataset. Combining SMOTE with machine learning algorithms such as Random Forests or Gradient Boosting Machines has proven effective in improving churn prediction accuracy by providing a more balanced training set.

In conclusion, existing research highlights the evolution of machine learning techniques in predicting customer churn, with each method offering unique strengths and limitations. While Logistic Regression provides simplicity and interpretability, advanced methods such as Random Forests, Gradient Boosting Machines, and Support Vector Machines excel in handling non-linear relationships and improving predictive accuracy. Furthermore, addressing the inherent imbalance in churn datasets through techniques like SMOTE has significantly enhanced the effectiveness of these models. These advancements underscore the importance of selecting and tuning the right combination of techniques to achieve robust churn prediction in the telecommunications industry.

CHAPTER 4

DATASETS

4.1 Datasets Description

This project utilizes the Telecom Customer Churn Dataset, Customer Churn Dataset, and Telecom Customer Usage Dataset sourced from Kaggle, a prominent platform for sharing and analyzing datasets. Published by IBM, this dataset provides comprehensive information on a hypothetical telecommunications company serving 7,043 customers in California during the third quarter of a fiscal year. The dataset is available in CSV format and includes detailed customer information for churn prediction analysis. The dataset contains the following key columns as following:

4.1.1 Customer Information

- customer_id, telecom_partner, gender, age, state, city, pincode, date_of_registration
- Additional attributes such as SeniorCitizen, Partner, and Dependents

4.1.2 Service Usage and Behavioral Data

- calls_made, sms_sent, data_used, Call Failure, Complains
- Seconds of Use, Frequency of use, Frequency of SMS, Distinct Called Numbers

4.1.3 Financial and Subscription Details

- Subscription Length, Charge Amount, Customer Value, MonthlyCharges, TotalCharges
- Tariff Plan, PaymentMethod

4.1.4 Churn Indicator

- Binary variable (Churn) indicating whether a customer left the service.

4.1.5 Additional Service Details

- PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies

4.1.6 Contract and Billing Information

- Contract, PaperlessBilling, Subscription Length

The dataset provides a wide range of features relevant for understanding customer behavior, service usage, and potential churn drivers. It includes demographic attributes (Age Group, gender), usage patterns, and churn-related information, making it highly suitable for developing predictive models.

The selection of this dataset is strategic, as it has been widely used in the data science community for telecom churn prediction projects. This enables comparisons with existing models and benchmarks, allowing us to evaluate the performance improvements of our model effectively. By leveraging this rich dataset, the project aims to develop a robust churn prediction model that provides actionable insights to improve customer retention strategies in the telecommunications industry.

CHAPTER 5

METHODOLOGY

The methodology for this project encompasses exploratory data analysis (EDA), preprocessing, machine learning model and Pipeline with a focus on aligning predictive outcomes with business goals. The steps below outline the approach taken to ensure the effective development of a churn prediction model.

5.1 Exploratory Data Analysis (EDA)

In this dataset, most of the features are categorical, which will need to be encoded. On the other hand, there are only 3 numerical features, which are tenure, monthly charges, and total charges. When we check the missing values, most of the features are free from missing values except the feature, total charges, which has some blank values causing this feature the data type of object although it should be float data type.

Then, we examine the general distributions of the important features and our target by plotting some histograms and boxplots. As expected, the class imbalance is found in the label column with 5174 values of no-churn customers and 1869 values of churn customers as shown in figures below

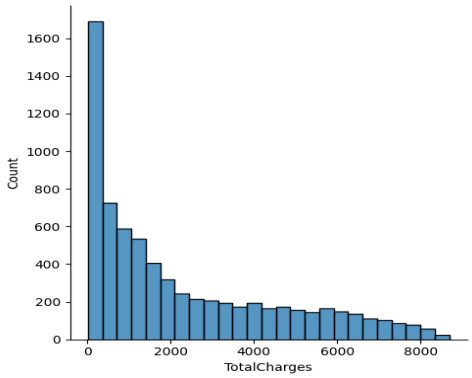

Figure 1. Distribution of Total Charge

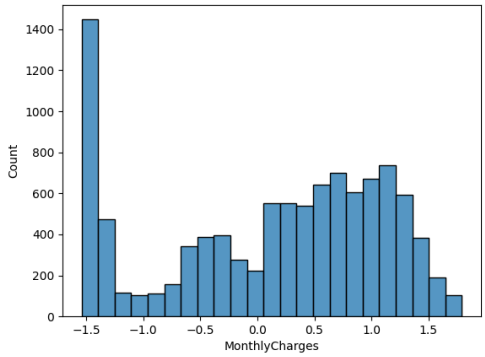

Figure 2. Distribution of Monthly Charges

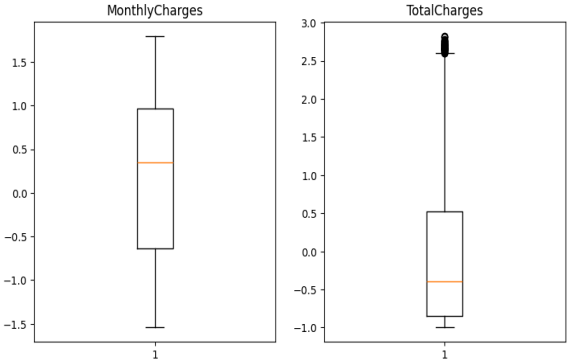

Figure 3. Box plots of Monthly Charges and Total Charges to see outliers

As mentioned, since we will use business-oriented measurements derived from the total revenue obtained from the customers to evaluate the model, we examine the percentage of the total charges that are above the median (1394.55). In Fig. 4, it can be seen that over 80 percent of the total revenue comes from customers with total charges above the median.

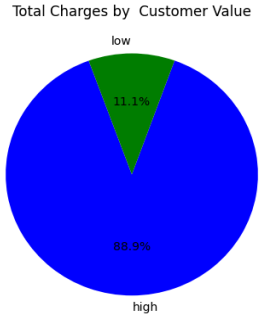

Figure 4. Percentage of Total Charges by high-value and low-value customers

Finally, we check the correlations and predictive power score, shown in Figure. 5 and 6, by plotting the heat maps to select the features. We have decided to exclude the tenure feature as it is highly correlated with some other features. The others are preserved.

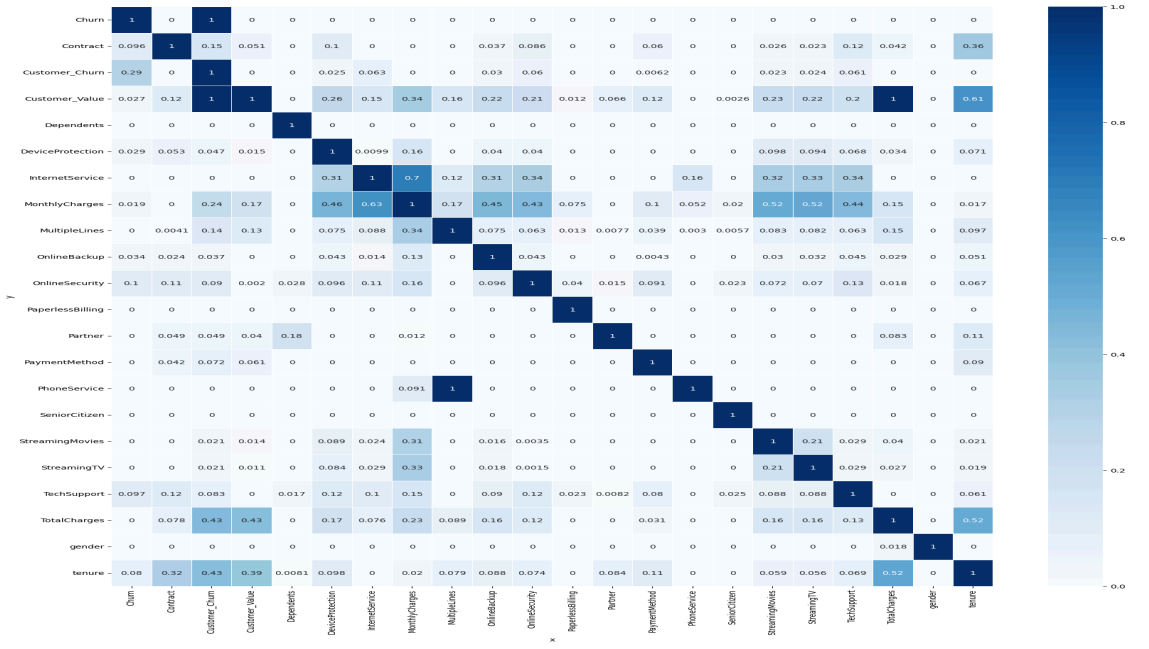

Figure 5. Predictive power scores

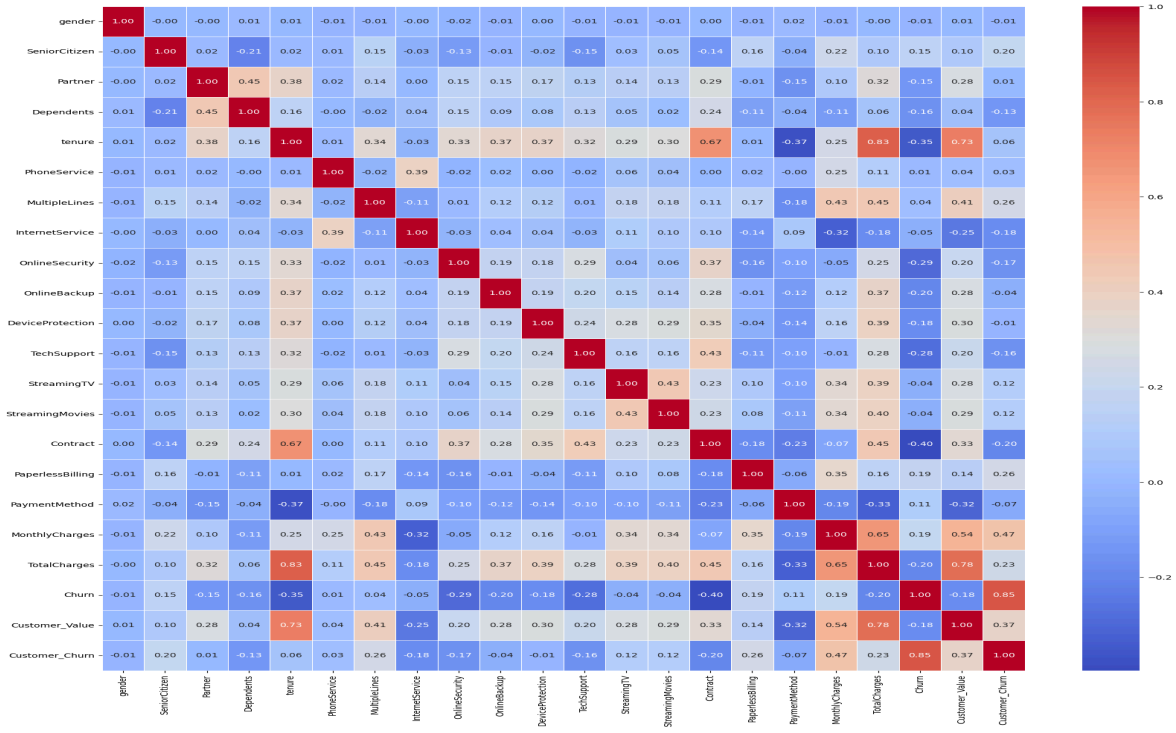

Figure 6. Correlation

5.2 Pre-Processing

After a thorough analysis on the dataset, we proceed to data preprocessing in order to perform necessary data manipulation actions so that the data are ready to feed the model. The following steps are performed.

5.2.1 Imputing missing values

To impute the missing data of the feature of total charges, we use the value in the monthly charge feature by assuming that these customers are and are not charged yet. Others are free from missing values.

5.2.2 Feature Engineering

we have created two new columns for model evaluation purposes as follows

1. Customer_value – “high” if the value is above the median and “low” if it is under the median value

2. Customer_churn – based on the Custmer_value and churn features, low_churn, high_churn, low_no_churn, and high_no_churn (Figure. 7)

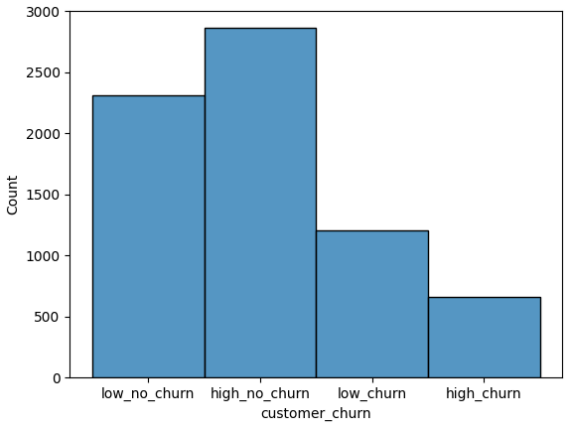

Figure 7. Distribution of four classes of new target

5.2.3 Split train-test

We have decided to use 10 percent of the data as the test set as around 800 samples are assumed to be enough for testing the model.

5.2.4 Scaling

We have decided to use StandardScaler() to scale the numerical features as they are charges that are not bounded in a certain range although the distributions are right-skewed.

5.2.5 Encoding

As most of the categorical features have only two to three unique values, we choose label encoding over one-hot encoding.

5.3 Machine learning model

In the preliminary step of model selection, we find the best among the following models by cross-validating the training set with five folds, by setting customer_churn columns with four classes as a target, and by setting weighted_recall as a scoring metric due to the class imbalance.

1. Logistic Regression
2. Support Vector Classifier
3. Decision Tree Classifier
4. Random Forest Classifier
5. AdaBoost Classifier
6. Gradient Boosting Classifier

Then, we select the gradient-boosting classifier as it has the highest mean weighted recall with around 7.9. To fine-tune the hyperparameters, the grid search is performed on this model. This model already outperforms the baseline model from Kaggle as shown in Figure 8 and Figure 9.

However, the model evaluation is purely done by comparing the weighted recall. The model from Kaggle has a weighted recall score of 7.76 and our model has 0.81. Due to the imbalanced dataset, recall is our main evaluator.

After the preliminary model selection, we upsample the data with SMOTE to enhance the performance. Then, we train a gradient-boosting classifier model with the target having two classes. We use this model as the baseline for evaluating the models with business-oriented measurements.

Gradient Boosting and histogram-based Gradient Boosting outperform other models. Therefore, we have tuned the hyperparameters of these two models using grid search. Here, in order to reduce the computational resources, we have tuned one hyperparameter at a time while others are left out as default values. The results are shown in Figure 10.

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.77	0.83	1035
1	0.54	0.76	0.63	374
accuracy			0.76	1409
macro avg	0.72	0.76	0.73	1409
weighted avg	0.80	0.76	0.78	1409

Figure 8. Classification report of the baseline model from Kaggle

	precision	recall	f1-score	support
0	0.81	0.89	0.85	223
1	0.82	0.99	0.90	288
2	0.77	0.63	0.69	130
3	0.40	0.03	0.06	64
accuracy			0.81	705
macro avg	0.70	0.64	0.62	705
weighted avg	0.77	0.81	0.77	705

Figure 9. Classification report of the best model from grid search on gradient boosting model

```
# gradeint boosting parameters
gbc_parameters = {'loss': 'log_loss',
'learning_rate': 0.1,
'min_samples_split': 0.17272727272727273,
'min_samples_leaf': 0.17272727272727273,
'max_depth': 3,
'max_features': 'log2',
'criterion': 'friedman_mse',
'subsample': 0.95,
'n_estimators': 50}

# histogram-based gradeint boosting parameters
hgc_parameters = {'learning_rate': 0.1,
    'min_samples_leaf': 10,
    'max_depth': None,
    'max_features': 0.5,
    'l2_regularization': 0.1,
    'max_iter': 130}
```

Figure 10. Hyperparameter tuning results

```
cls_weights1 = {0 : 2, 1 : 1, 2 : 1, 3 : 1}
cls_weights2 = {0 : 3, 1 : 1, 2 : 2, 3 : 1}
cls_weights3 = {0 : 4, 1 : 1, 2 : 1, 3 : 1}
cls_weights4 = {0 : 4, 1 : 1, 2 : 1, 3 : 1}
```

Figure 11. Classification weights setting of histogram-based gradient-boosting classifier

Finally, we train a gradient-boosting classifier and a histogram-based gradient-boosting classifier, which enable us to define the class weight, with the target having four classes. In the

histogram-based gradient-boosting classifier, we define the class weights of each class for four experiments, shown in figure 11, where 0 = high_churn and 2 = low_churn and others are no_churn.

5.4 Machine learning Pipeline

The machine learning pipeline consists of two parts, preprocessing and classification. The preprocessing pipeline consists of imputation and scaling for numerical features and encoding for categorical features. Based on the EDA results, we choose Standard Scaler for scaling and Ordinal Encoder for encoding. Then, we use a simple imputer for missing values, by passing the mean for numerical features and mode for categorical features into the strategy parameter. These preprocessing steps are combined with a column transformer.

This preprocessing pipeline is concatenated with the classifier, forming the complete machine learning pipeline, for training, testing and serving. In case of parameter tuning, this pipeline is passed to grid search. The overview can be seen in figure 12.

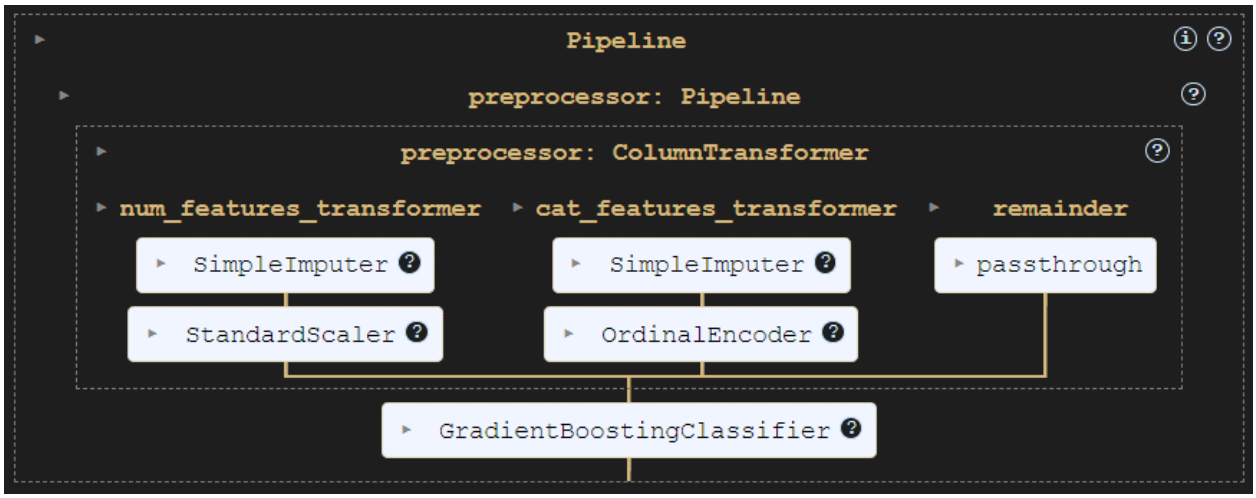

Figure 12. Machine Learning Pipeline Structure

CHAPTER 6

MODEL EVALUATION RESULTS

We will evaluate the models by measuring how much the model can potentially cause revenue losses. In this context, the false negatives are the main concern as the company is not aware of potential customer churn although these customers have the potential to churn. Moreover, the impact on revenue is higher if the high-value customers churn.

Hence, we will use the sum of the total charges of the false negatives instead of simple false negative counts. Therefore, the risk factor which is measured by the potential loss of revenue due to false negatives predicted by the model will be the main indicator of the model performance. It can be summarized in the equation (1).

$$Potential\ Risk\ in\ Currency\ = \sum_{i\,=\,0}^{m} (TC)\ if\ FN \qquad (1)$$

$$Risk\ Factor\ = \frac{Potential\ Risk\ in\ Currency}{\sum_{i\,=\,0}^{m} (TC)\ where\ target = "churn"} \qquad (2)$$

where,

TC is the total charge,

FN is a false negative predicted by the model,

m is the number of observations,

The primary goal of the model evaluation process is to assess how well each model minimizes potential revenue losses caused by **false negatives**. In the context of customer churn, **false negatives**—customers incorrectly predicted as non-churners—pose a significant risk, as the company may fail to address their dissatisfaction and prevent them from leaving. This issue becomes even more critical when the false negatives include high-value customers, as their departure has a greater impact on overall revenue.

6.1 Evaluation Metrics

To accurately measure the financial implications of false negatives, we have adopted the following evaluation framework.

6.1.1 Potential Risk in Currency

This metric calculates the total potential revenue loss caused by false negatives. Instead of merely counting the number of false negatives, this approach sums the total charges associated with these misclassified customers, providing a clearer picture of the financial risk.

6.1.2 Risk Factor

The risk factor is a normalized metric derived from the revenue loss due to false negatives relative to the total revenue of actual customer churns. It highlights the proportion of potential revenue loss caused by the model's misclassifications.

6.1.3 Number of Wrong Predictions

While not the primary focus, the total number of wrong predictions, including both false positives and false negatives, provides additional context for overall model performance.

6.1.4 False Negatives

The count of false negatives serves as a secondary metric to understand the scale of misclassification. However, the focus remains on the financial impact of these false negatives.

	Models	Potential Risk in Currency	Risk Factor	Number of wrong predictions	False Negatives
0	gbc_4class_upsample	168376.30	0.520904	191	64
1	hgbc_4class_upsample_class_weight4	185005.20	0.572349	190	77
2	hgbc_4class_upsample_class_weight3	186503.45	0.576984	193	69
3	hgbc_4class_upsample_class_weight1	187950.35	0.581460	168	76
4	hgbc_4class_upsample_class_weight2	188011.10	0.581648	197	65
5	gbc_2class_upsample	190178.70	0.588354	178	63
6	hgbc_4class_upsample	236796.05	0.732573	167	90
7	hgbc_4class	246049.50	0.761201	157	89
8	gbc_2class	247608.90	0.766025	147	100
9	gbc_4class	277095.30	0.857247	145	105

Table 1. Evaluation Results

From table1, the summarizes of evaluation results for various models, including variations of Gradient Boosting Classifier (GBC) and Histogram-Based Gradient Boosting Classifier (HGBC). The results show the potential risk in currency, risk factor, and false negatives for each model configuration.

1. Models with Upsampling: Models utilizing upsampling techniques generally perform better in reducing potential revenue loss. For example, gbc_4class_upsample has a significantly lower risk factor (0.5209) compared to gbc_4class (0.8572), indicating its ability to better identify high-value customers at risk of churn.

2. Class Weight Adjustments: Fine-tuning class weights further improves the model's ability to prioritize high-value churners. For instance, hgbc_4class_upsample_class_weight1 achieves a lower potential risk (187950.34) and risk factor (0.5814) compared to the unweighted version, hgbc_4class_upsample.

3. Trade-offs: Some models, such as gbc_2class, exhibit higher false negatives and risk factors (0.7660), emphasizing the importance of balancing accuracy with financial impact.

The evaluation highlights that the most effective models are those that incorporate upsampling and class-weight adjustments. These methods significantly reduce the financial risks posed by false negatives, particularly for high-value customers. By prioritizing potential revenue loss as the primary indicator of performance, the evaluation aligns directly with the business objective of maximizing profitability while addressing class imbalance effectively.

CHAPTER 7

DISCUSSION

As the customer churn prediction is well-studied, there are many different models developed. As mentioned in the modeling section, a model from Kaggle is used as a baseline performance. In the modeling selection, we have already developed a gradient boosting model that has more accuracy than the baseline performance. As our objective is to enhance the model performance in terms of business oriented metrics, which we have introduced in the equation (1) and (2), we further develop a model with various techniques. This includes framing binary classification problem to multi nominal classification problem to evaluate the quality of prediction, upsampling with SMOTE, parameter tuning with grid search and setting the class weights.

Through these techniques, our model is much improved in terms of the main evaluation metric, risk factor although this model might be lower in other metrics based on the counts, such as accuracy, precision and recall. As the risk factor is an indicator of how well the model predicts the high-valued customer churns, our model has much more the potential to save revenue loss than others which are merely focused on the measured counts. The evaluation results show that the model with upsampling of four target classes has the risk factor of around 0.5 while it is around 0.7 for the two classes model without upsampling. Therefore, we can conclude that our model achieves our aims.

On the other hand, the model complexity and interpretability are fairly the same across different models as the hyperparameter setting is the same for all models.

CHAPTER 8

CONCLUSION

In this project, we have introduced an evaluation metric that can be used to gauge the risk of the model in terms of potential revenue loss. This metric is suitable for the customer prediction problem, in which the data is generally imbalanced and the least dominant class has higher interest, such as revenue. This metric can be used for other problems which have the same properties.

Moreover, we have demonstrated that recategorizing the target classes based on the objective of problems and then upsampling the data if the interested class in the target is suffering from the class imbalance is a suitable technique to achieve the objective. In short, we have delivered a more robust model for a specific business use case using these techniques which can be utilized in other problems with the same scenario.

REFERENCES

1. Wanchai, P. (2017). Customer churn analysis: A case study on the telecommunication industry of Thailand. 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), Cambridge, UK, 325–331. https://doi.org/10.23919/ICITST.2017.8356410
2. Parmar, M. (n.d.). Customer churn classification. Retrieved from https://www.kaggle.com/code/mehulparmar2712/customer-churn-classification
3. Tsai, C.-F., & Lu, Y.-C. (2009). Customer churn prediction with data mining techniques. Expert Systems with Applications, 36 (10), 12581–12589. https://doi.org/10.1016/j.eswa.2009.03.015
4. Idris, A., Chuan, L. S., & Rahman, A. A. (2012). Random forests for customer churn prediction. International Journal of Computer Applications, 53 (10), 1–7. https://doi.org/10.5120/8686-5303
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
6. Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing with logistic regression. Computers & Operations Research, 35 (10), 3283–3292. https://doi.org/10.1016/j.cor.2007.06.014
7. Chawla, N. V., De'Belle, M. L., & Lazarevic, A. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953