Proposal: Telcom Customer Churn Prediction with Machine Learning

\*For the fulfillment project proposal of AT82.01 Computer Programming for Data Science and Artificial Intelligence course by Dr. Chantri Polprasert

1st Kaung Nyo Lwin

Department of Data Sciences and Artificial Intelligence

Institute of Technology

Pathum Thani, Thailand

st125066@ait.asia

**Abstract—The project proposal is to predict customer churn, enhancing the impact on business value, based on a Kaggle public dataset.**

1. *INTRODUCTION*
2. Introduction to Telecommunication customer churn

In the highly competitive telecommunication industry, customer retention has become increasingly vital for sustaining business growth and profitability. Customer churn, or the rate at which users discontinue their subscriptions, poses a significant threat to revenue, as acquiring new customers often costs more than retaining existing ones. Additionally, high churn rates may signal underlying issues such as poor service quality, customer dissatisfaction, or competitive pressure, making it critical for telecom providers to proactively manage churn.

1. Telecommunication Business Perspective on Customer Churn

From a business perspective, reducing customer churn is crucial for maintaining profitability, as retaining existing customers is significantly more cost-effective than acquiring new ones. Churn reduction has a direct impact on Customer Lifetime Value (CLTV), a key metric for long-term business growth, as customers who stay with the

2nd  Patsachon Pattakulpong

Department of Data Sciences and Artificial Intelligence

Institute of Technology

Pathum Thani, Thailand

st124952@ait.asia

company generate higher revenue over time. Additionally, high churn rates often signal poor customer satisfaction, which can damage brand loyalty. By focusing on management, telecom companies can enhance their brand reputation, foster customer loyalty, and encourage positive word-of-mouth. In an increasingly competitive market, effective churn management is essential to prevent customers from switching to competitors who may offer better pricing or services, thus maintaining a strong market position.

1. Why are we interested in telecommunications customer churn projects?

1. Proactive Customer Retention: Churn prediction enables companies to identify at-risk customers early, allowing for targeted retention strategies. This proactive approach helps to minimize customer loss before it occurs.

2. Cost Efficiency: Retaining existing customers is significantly more cost-effective than acquiring new ones. By reducing churn, telecom companies can lower marketing and acquisition costs, leading to improved profitability.

3. Increased Customer Lifetime Value (CLTV): Churn prediction contributes to maximizing Customer Lifetime Value. Retained customers generate more revenue over time, enhancing long-term business growth and stability.

4. Enhanced Customer Experience: Understanding the factors that lead to churn allows telecom companies to improve their services and tailor offerings to meet customer needs. This leads to higher satisfaction and loyalty, further reducing churn.

*II. PROBLEM STATEMENT*

In the highly competitive telecommunications industry, customer churn poses a significant challenge that directly impacts revenue and profitability. As customer loyalty decreases and switching costs become minimal, telecom companies are experiencing rising churn rates, resulting in substantial financial losses and increased customer acquisition costs. Despite investing in marketing and customer service initiatives, the inability to effectively predict and understand the underlying factors contributing to customer churn limits the capacity for proactive retention strategies.

Our project aims to develop a robust predictive model that accurately identifies at-risk customers by analyzing various factors, including usage patterns, customer demographics, service quality, and historical churn data. By leveraging machine learning techniques, the objective is to enhance the company's ability to implement targeted retention efforts, optimize marketing resources, and ultimately reduce churn rates, leading to improved customer satisfaction and increased long-term revenue.

*III. RELATED WORKS*

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.

Logistic Regression is a commonly used technique in churn prediction due to its simplicity and interpretability. Studies have shown that it is effective when baseline accuracy is needed. Researchers like **Tsai and Lu (2009)** used Logistic Regression to model customer churn, focusing on feature importance to understand the key drivers of churn, such as contract duration, data usage, and service complaints. However, this method struggles with more complex, non-linear relationships in customer behavior.

Decision Trees and Random Forests have been widely adopted for telecom churn prediction due to their ability to handle non-linear relationships and provide interpretable models. **Idris et al. (2012)** applied Random Forests to customer churn prediction, achieving better accuracy than logistic models. By utilizing multiple decision trees, Random Forests help reduce the overfitting problem associated with single decision trees and are effective in managing large, unbalanced datasets.

**Gradient Boosting Machines (GBM)** are widely recognized for their superior performance in customer churn prediction tasks. **Chen and Guestrin (2016)** used XGBoost, a highly optimized implementation of GBM, to predict customer churn. Their approach focuses on iteratively improving weaker models, which results in high predictive accuracy. GBM methods have consistently been top performers in telecom churn competitions and real-world applications, with the trade-off of requiring more computational power and tuning. **Coussement and Van den Poel (2008)** applied **Support Vector Machines (SVM)** to telecom churn prediction and reported that SVM outperformed traditional methods like logistic regression and decision trees. SVM can model complex, non-linear decision boundaries between churn and non-churn customers by using kernel methods. However, SVM requires careful tuning of hyperparameters (such as the regularization parameter and kernel function) and can be computationally intensive for large datasets.

Customer churn datasets are typically imbalanced, meaning there are fewer churned customers compared to non-churned ones. To address this, **Chawla et al. (2002)** introduced the **SMOTE (Synthetic Minority Over-sampling Technique)** to balance datasets by creating synthetic samples of the minority class. This technique, combined with machine learning algorithms like Random Forests or Gradient Boosting, can significantly improve churn prediction accuracy by providing the model with a more balanced training set.

*IV. DATASET*

1. Description

This dataset is sourced from Kaggle, a public platform for sharing datasets and data analysis projects. The Telco customer churn dataset encompasses details regarding a hypothetical telco company catering to 7043 customers in California during quarter 3 of the fiscal year, published by IBM. This dataset is used in this project as many models developed on Kaggle using this dataset to compare our results and illustrate the improvements of our model. The dataset is in CSV format.

1. Features

The dataset contains details on customers who recently left, the services they have subscribed to, as well as their account information such as tenure length, contract type, payment method, and billing preferences. Additionally, it includes demographic information like gender, age range, and whether they have partners and dependents.

*V. METHODOLOGY*

Conventionally, model performance is assessed using metrics like accuracy, precision, and recall. However, for this project, we will prioritize evaluating the models based on business-oriented measurements derived from the total revenue obtained from customers. This approach will enable us to maximize the profitability of the company while effectively addressing class imbalance Our toolkit will include pandas, matplotlib, Seaborn for exploratory data analysis, SMOTE for data resampling, and scikit-learn for modeling.

1. Data Acquisition

This dataset is publicly available on Kaggle. We download it from the link <https://www.kaggle.com/datasets/blastchar/telco-customer-churn/data>.

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 three 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 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 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. 7, it can be seen that over 80 percent of the total revenue comes from customers with total charges above the median.

Finally, we check the correlations and predictive power score 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.

1. 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.

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.

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 unde 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

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.

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.

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

1. Modeling

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 Classifer
      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. However, the model evaluation is purely done by comparing the weighted recall. This can be seen in Fig. 2 and Fig. 3.

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.

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 weight of the high\_churn as 2 and 1 for others to address the importance of this class.

1. Evaluation

We will evaluate the models by measuring how much the model potentially can 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.

where,

TC is the total charge,

FN is a false negative predicted by the model

1. Deployment

We will implement a dash web application for hosting. Then, we will deploy our model to the dash app. After testing in the local host, we will upload the app files to GitHub along with the docker file for live server deployment. We intend to use AWS to host our final deliverable.

*VI. PRELIMINARY RESULTS*

After evaluating on the last three models, the results are summarized in the Fig. 1. It can be seen that the models trained on the target have four classes have a considerable amount of reduced revenue loss caused by false negatives.

On the other hand, the weighted average recall scores for the three models in Fig.1 are not much different from each other. In fact, Histogram-based XGboost has even less weighted recall scores compared to other twos. This is because these scores are just the counts without taking into account the impact of false negatives. Howerver, in terms of business values, this lowest-scoring model is the best as it minimizes the revenue loss.

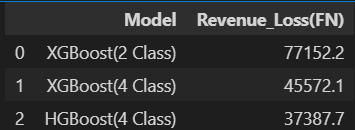
**

Fig. 1: Potential revenue Loss due to false negatives predicted by the models

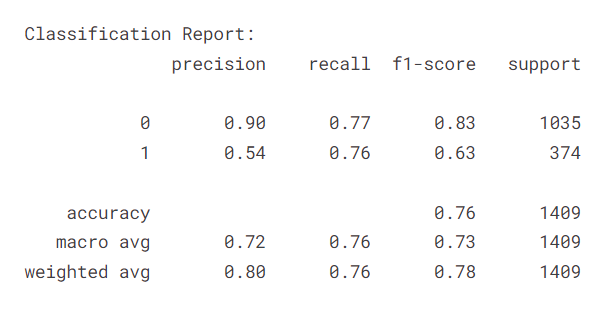


Fig. 2: Scores of the baseline model (Kaggle)

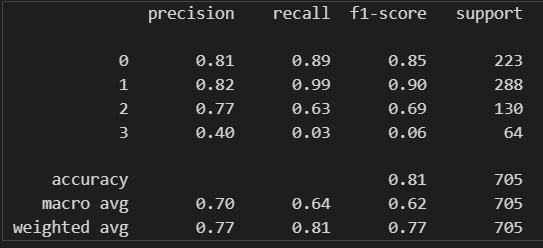


Fig. 3: Scores of grid search on XGboost

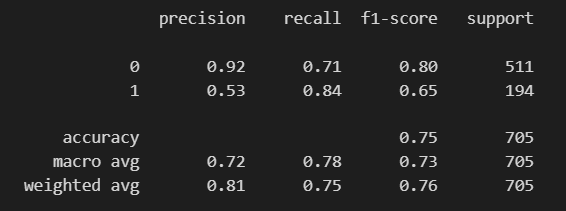


Fig. 4: Scores of XGboost with two classes

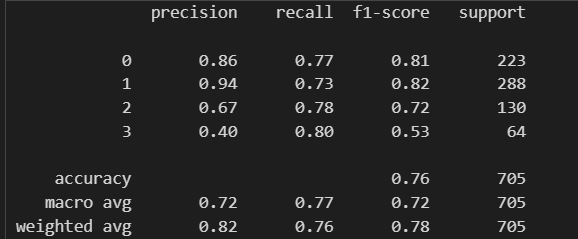


Fig. 5: Scores of XGboost with four classes

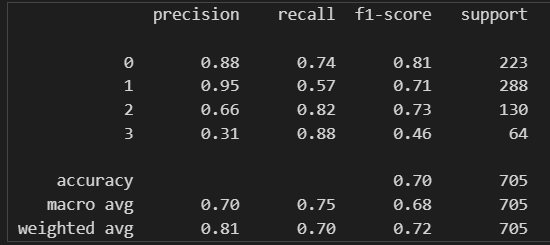


Fig. 5: Scores of Histogram-based XGboost with four classes by setting class weight 2 to high\_churn class

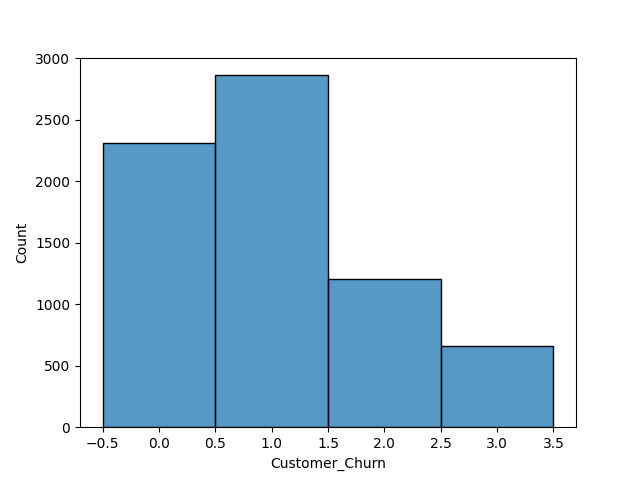


Fig. 6: Histogram of Customer churn four classes

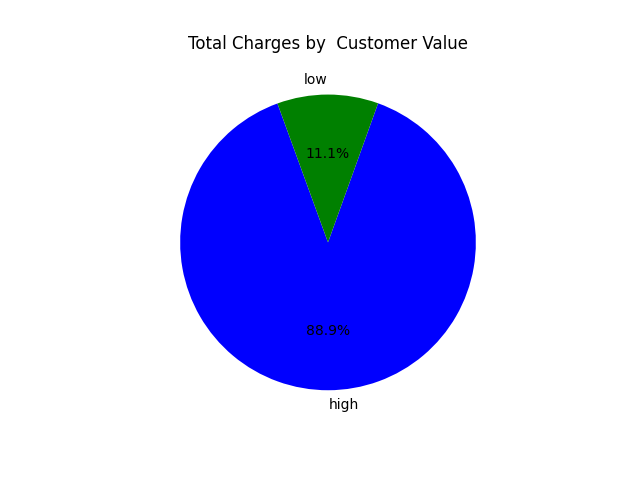
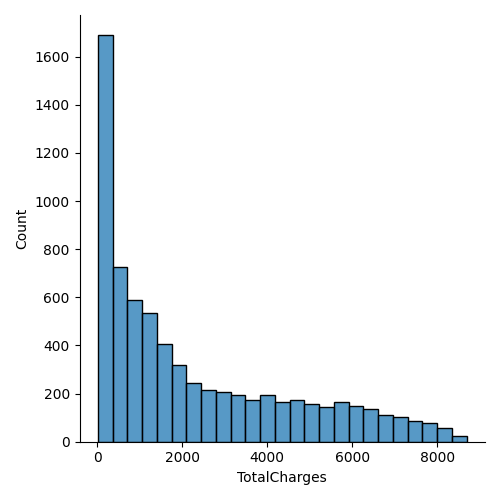


Fig. 7: Percentage of total charges by customer value

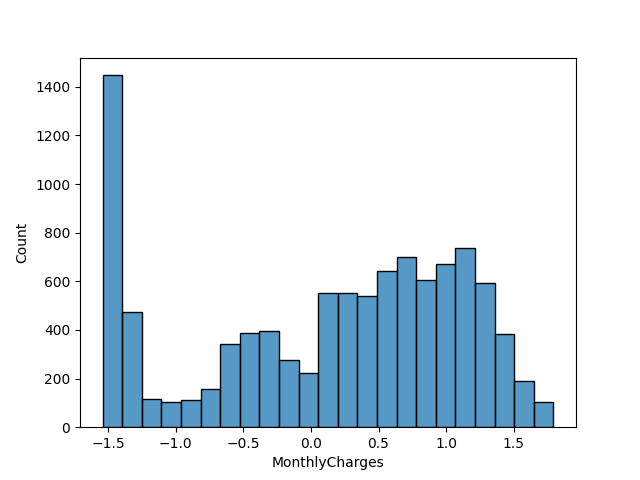
References

[1] P. Wanchai, "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, 2017, pp. 325-331, doi: 10.23919/ICITST.2017.8356410. keywords: {Data mining;Companies;Communications technology;Decision trees;Predictive models;Data models;Industries;component;churn prediction;telecommunication industry;data mining;CRM;decision trees},

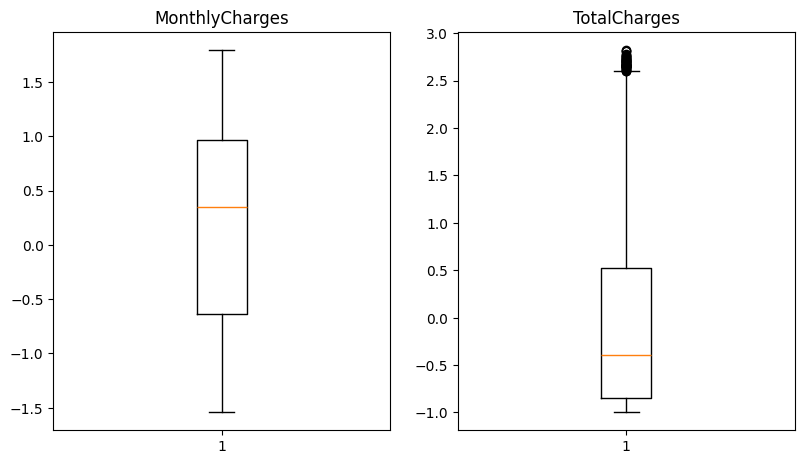
[2] https://www.kaggle.com/code/mehulparmar2712/customer-churn-classification



(a) Total Charge

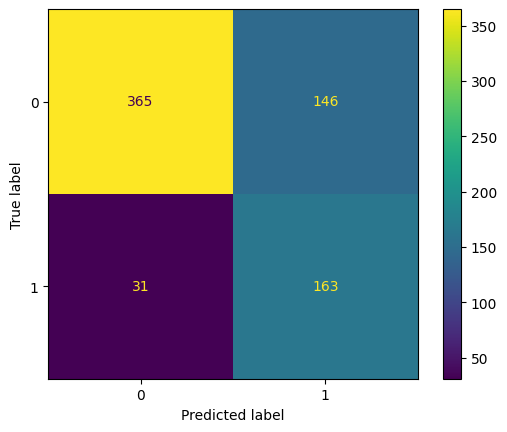


(b) Monthly Charges

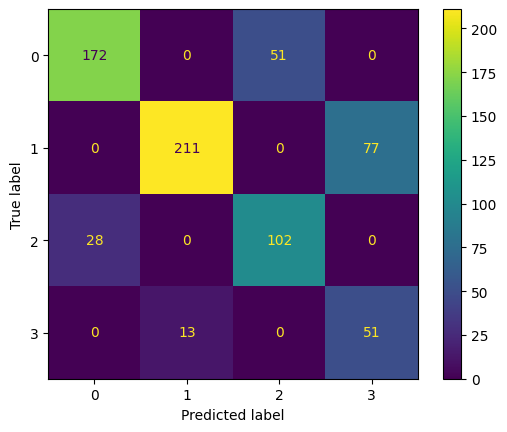


(c) Box plots

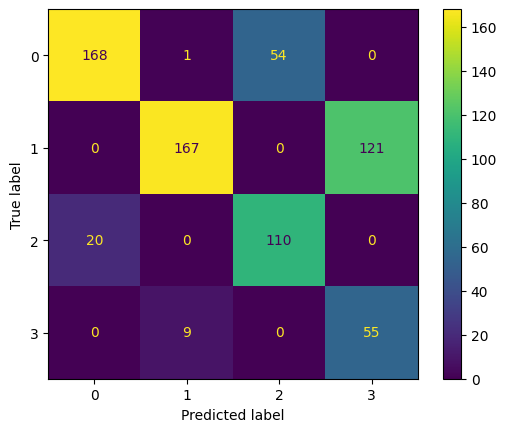
Fig. 8: Distributions of Total charges and monthly charges



(a) XGBoost with two classes



(b) XGBoost with four classes



(c) Histogram-based XGboost with four classes

Fig. 10: Confusion Matrix of models

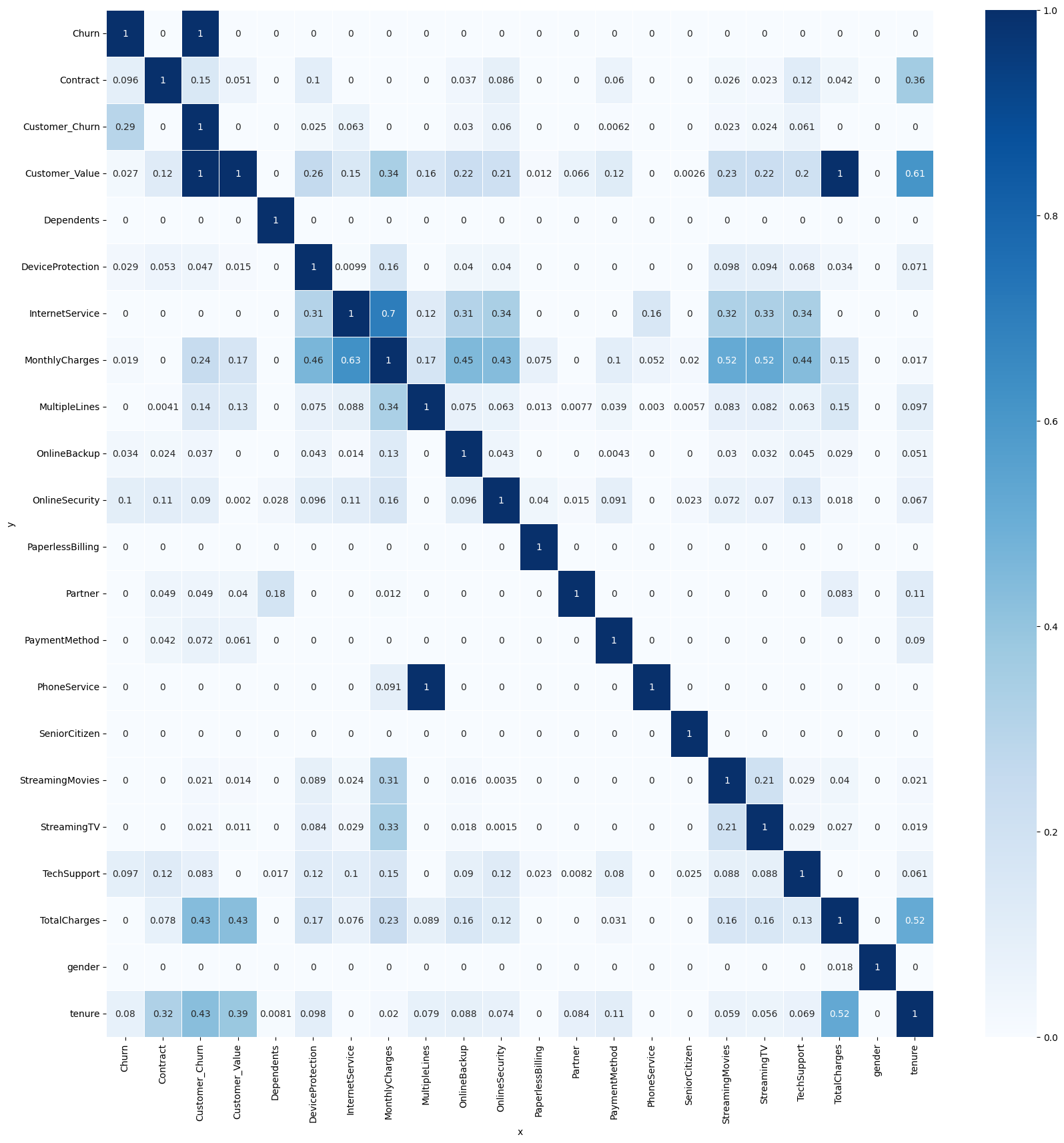


Fig. 9: Predictive power scores

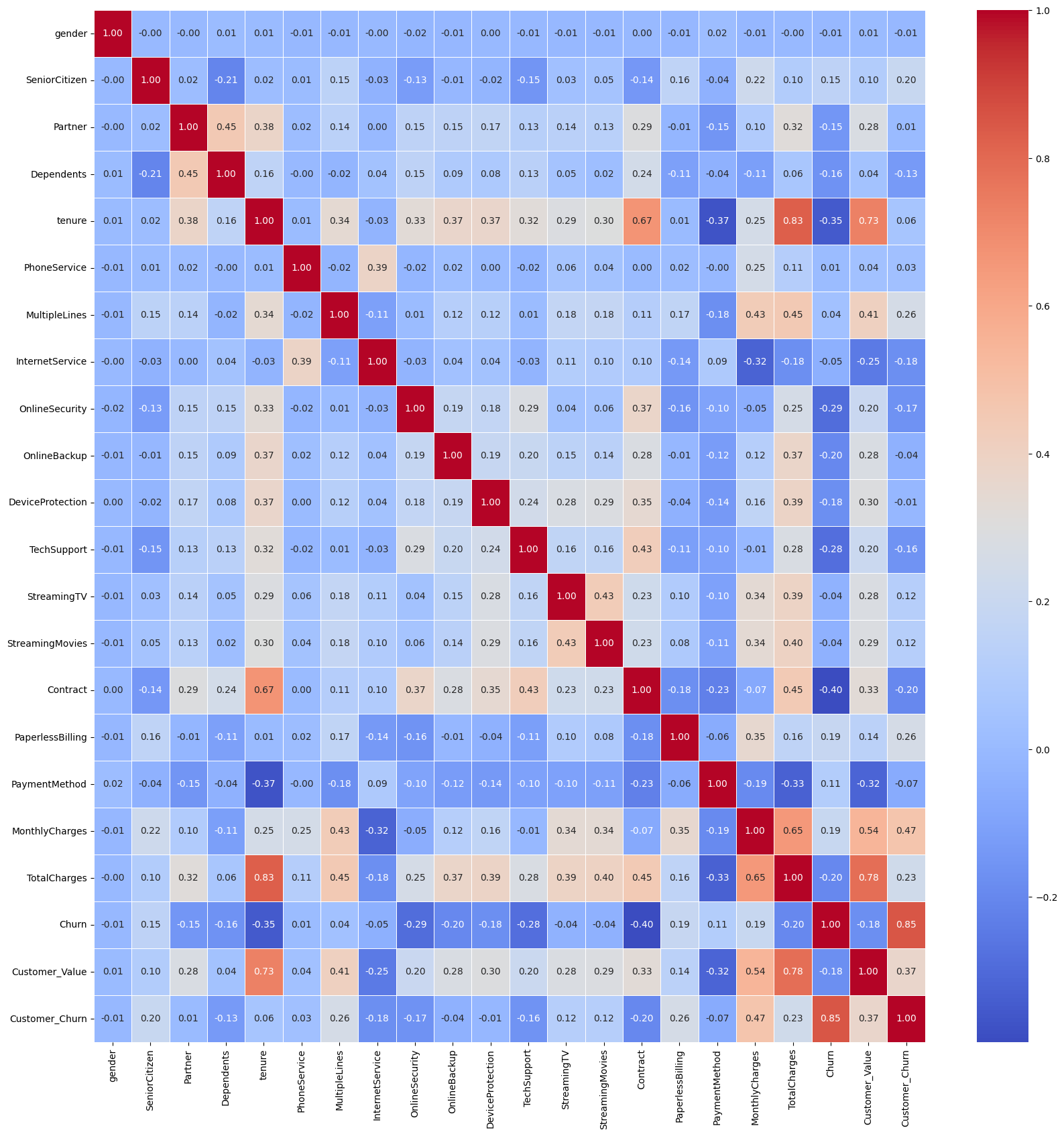


Fig. 9: Correlation matrix