Turcin Cristian
Deaconu Șerban

Report for Customer Churn Prediction

1. Introduction

Customer churn represents the phenomenon in which customers stop using a product or service. In competitive business environments, retaining existing customers is crucial, as acquiring new customers usually involves higher costs than maintaining current ones. Consequently, identifying customers who are likely to churn has become an important objective for many organizations.

The ability to accurately predict customer churn allows companies to take proactive measures, such as personalized offers or targeted interventions, in order to reduce customer loss. As a result, churn prediction has become a common application of machine learning techniques in business analytics.

From a machine learning perspective, this task represents a supervised binary classification problem, where the goal is to predict whether a customer will churn or not based on a set of descriptive features. The objective of this project is to build and compare multiple machine learning models for customer churn prediction. Different preprocessing techniques are applied, and several classification algorithms are evaluated in order to determine the most effective approach for this task.

2. Data

This section describes the dataset used in the project, as well as the data cleaning and preprocessing steps applied before model training.

2.1. Dataset Description

The dataset used in this project contains information about customers of a subscription-based service. Each instance in the dataset represents a single customer and includes a variety of features describing demographic characteristics, service usage, customer engagement, and financial information.

The target variable is churn, a binary variable that indicates whether a customer has unsubscribed from the service. A value of 1 corresponds to a churned customer, while a value of 0 represents a non-churned customer.

The dataset consists of approximately 10,000 records and multiple input features. An analysis of the target variable distribution shows that the dataset is imbalanced, with a significantly smaller proportion of churned customers compared to non-churned customers. This imbalance poses additional challenges for classification and influences the choice of evaluation metrics used in this project.

Starting from the original dataset, several dataset variants were created in order to analyze the impact of feature selection and dimensionality reduction on model performance. These variants include a dataset with a large number of features, a cleaned dataset containing only relevant features, a reduced dataset based on qualitative feature selection, and a dataset obtained through variance-based dimensionality reduction using Principal Component Analysis (PCA).

2.2. Data Cleaning

In order to prepare the data for analysis and modeling, several data cleaning strategies were applied, resulting in multiple dataset variants derived from the original dataset. Each variant was created with a specific purpose in mind, allowing the impact of feature selection on model performance to be evaluated.

2.2.1. Many Features Dataset

For the first dataset variant, referred to as the many-features dataset, most of the original features were retained in order to preserve as much information as possible. After removing identifier-type attributes and certain marketing and support-related features that do not provide meaningful predictive value for customer churn, the resulting dataset contains 27 features. The excluded features include the customer id, email open rate, marketing click rate, and complaint type.

2.2.2. Relevant Features Dataset

The second dataset variant, referred to as the relevant-features dataset, was obtained by applying a more aggressive feature selection strategy based on domain knowledge. In this case, features considered irrelevant or weakly related to churn prediction were removed, resulting in a dataset containing 11 features. These removed features include demographic attributes, customer profile information, usage and engagement metrics, pricing and billing details, as well as marketing and support-related features. The goal of this step was to retain only variables that are more likely to contribute directly to predicting customer churn.

2.2.3. High Quality Features Dataset

A third dataset variant was constructed using qualitative feature selection, resulting in a high-quality-features dataset composed of only two variables: the Customer Satisfaction Score (CSAT) and the Net Promoter Score (NPS). These features were selected based on their strong conceptual relationship with customer churn, as customer satisfaction and willingness to recommend a service are widely recognized indicators of customer retention.

2.2.4. PCA Variant

The final dataset variant, which focuses on dimensionality reduction, is created later in the preprocessing stage using Principal Component Analysis (PCA). This approach generates a reduced representation of the data based on variance, allowing for further analysis of the effect of feature reduction on model performance.

2.3. Preprocessing

2.3.1. Encoding

Several features in the dataset are categorical and represented as string values. Since machine learning algorithms such as Logistic Regression and Support Vector Machines require numerical input, these categorical variables must be transformed into a numerical format.

As the categorical features are nominal and do not represent an inherent numerical order, One-Hot Encoding was applied. This encoding technique prevents the introduction of artificial ordinal relationships, which could negatively affect the performance of linear and distance-based models.

All categorical features were automatically identified based on their data type and encoded using One-Hot Encoding. Numerical features were left unchanged at this stage and were processed separately during the scaling step.

2.3.2. Splitting the Dataset

The dataset is split into training and test sets using a 70%–30% ratio. Stratified sampling is applied in order to preserve the original class distribution of the target variable in both subsets. A fixed random seed is used to ensure reproducibility of the results.

2.3.3. Scaling

Feature scaling is applied using the StandardScaler, which standardizes numerical features by removing the mean and scaling them to unit variance. This step is particularly important for classification models such as Logistic Regression and Support Vector Machines, which are sensitive to the scale of input features.

2.3.4. Dimensionality Reduction (PCA)

Principal Component Analysis (PCA) with three components is applied after feature scaling. The transformation is fitted on the training data only to avoid data leakage. PCA reduces the dimensionality of the feature space by projecting the data onto directions that capture the maximum variance, resulting in a compact representation of the original dataset. This reduced dataset is later used to analyze the impact of dimensionality reduction on model performance.

3. Algorithms

3.1. Logistic Regression

Logistic Regression is a supervised learning algorithm commonly used for binary classification problems. It models the probability of a given instance belonging to a specific class using a logistic (sigmoid) function.

One of the main advantages of Logistic Regression is its simplicity and interpretability. The model is computationally efficient and performs well when there is a linear relationship between the input features and the target variable. Additionally, it works effectively when combined with proper feature scaling.

However, Logistic Regression has limitations. As a linear model, it struggles to capture complex, non-linear patterns in the data. Its performance may degrade

when the relationship between features and the target variable is highly non-linear or when important interactions between features are present.

In this project, Logistic Regression was implemented with specific hyperparameters. The class_weight parameter was set to "balanced" to address the class imbalance present in the dataset by assigning higher importance to the minority class. A fixed random_state was used to ensure reproducibility of the results. The maximum number of iterations was increased to 1000 in order to allow the optimization algorithm to converge properly.

3.2. Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised learning algorithms used for classification tasks. SVM aims to find an optimal decision boundary that maximizes the margin between different classes. By using kernel functions, SVM can model complex non-linear relationships between features.

One of the main advantages of SVM is its ability to handle high-dimensional data and capture non-linear patterns effectively, especially when combined with kernel functions such as the Radial Basis Function (RBF). SVM models are robust to overfitting when properly regularized. However, SVMs can be computationally expensive for large datasets and require careful tuning of hyperparameters.

In this project, an SVM classifier with an RBF kernel was used. The class_weight parameter was set to "balanced" to compensate for the class imbalance present in the dataset. The regularization parameter C was set to 0.1, controlling the trade-off between maximizing the margin and minimizing classification error. The gamma parameter was set to "scale", which automatically adjusts the kernel coefficient based on the input features. A fixed random_state was used to ensure reproducibility of the results.

3.3. Decision Tree

Decision Trees are supervised learning algorithms that perform classification by recursively splitting the feature space based on decision rules. Each split is selected in order to maximize the homogeneity of the resulting subsets.

In this project, Decision Tree classifiers are trained using the Gini impurity criterion, which measures the quality of a split by evaluating how mixed the classes are within a node. One of the main advantages of Decision Trees is their interpretability, as the decision process can be easily visualized and understood. Additionally, tree-based models do not require feature scaling and are not affected by the scale of input features.

However, Decision Trees are prone to overfitting, especially when trained without constraints on tree depth or node size. Although dimensionality reduction is not required for this type of model, the classifier is also trained on the PCA-reduced dataset for comparison purposes, in order to analyze the impact of feature reduction on its performance.

In the implementation, class_weight was set to "balanced" to address class imbalance, and a fixed random_state was used to ensure reproducibility of the results.

4. Results

The performance of the models was evaluated using Accuracy and F1-score. While accuracy provides a general measure of correct predictions, the F1-score is more appropriate for imbalanced classification problems, as it balances precision and recall. Therefore, greater emphasis is placed on the F1-score when comparing model performance.

4.1. Logistic Regression Results

Dataset Variant	Accuracy	F1-Score
Many Features	0.675	0.285
Relevant Features	0.654	0.265
High-Quality Features	0.540	0.215

PCA(3 components)	0.544	0.208

Table 1. Logistic Regression performance across dataset variants

The Logistic Regression model achieved the best performance when trained on the dataset containing many features. A gradual decrease in F1-score can be observed as the feature space is reduced, with the lowest performance obtained on the high-quality and PCA-reduced datasets.

4.2. Support Vector Machine Results

Dataset Variant	**Accuracy**	**F1-Score**
Many Features	0.688	0.323
Relevant Features	0.697	0.305
High-Quality Features	0.816	0.278
PCA(3 components)	0.624	0.233

Table 2. Support Vector Machine performance across dataset variants

The best performance was obtained when using the dataset with many features, indicating that SVM benefits from a richer feature representation. Although the highest accuracy was observed on the high-quality dataset, the corresponding F1-score was lower, suggesting that the model struggled to correctly identify the minority class. Dimensionality reduction using PCA resulted in a noticeable decrease in performance.

4.3. Decision Tree Results

Dataset Variant	**Accuracy**	**F1-Score**

Many Features	0.829	0.163
Relevant Features	0.842	0.218
High-Quality Features	0.653	0.190
PCA(3 components)	0.810	0.136

Table 3. Decision Tree performance across dataset variants

The Decision Tree model achieved relatively high accuracy values across all dataset variants. However, its F1-scores remained consistently low, indicating poor performance in identifying churned customers. The best F1-score was obtained using the dataset with relevant features, while both the high-quality and PCA-reduced datasets resulted in reduced performance. These results suggest that, despite its interpretability, the Decision Tree model struggles with class imbalance and is sensitive to information loss.

4.4. Comparative Analysis

When comparing the three classification models, Support Vector Machines consistently achieved the highest F1-scores across all dataset variants. Logistic Regression showed moderate performance, while Decision Trees obtained the lowest F1-scores despite relatively high accuracy values. These results highlight the importance of selecting appropriate evaluation metrics and models when dealing with imbalanced datasets.

5. Conclusions

In this project, multiple machine learning models were developed and evaluated for the task of customer churn prediction. The experimental results show that the dataset is highly imbalanced, which significantly affects model performance and highlights the importance of using appropriate evaluation metrics such as the F1-score.

Among the evaluated algorithms, Support Vector Machines achieved the best overall performance, consistently obtaining the highest F1-scores across all

dataset variants. Logistic Regression provided moderate results, demonstrating reasonable performance while maintaining simplicity and interpretability. Decision Trees, despite achieving relatively high accuracy values, performed poorly in terms of F1-score, indicating difficulties in correctly identifying churned customers.

The experiments also revealed that feature representation plays a crucial role in churn prediction. Models trained on datasets with a larger number of features generally performed better, while dimensionality reduction using PCA led to decreased performance for all algorithms. This suggests that, for this dataset, preserving feature information is more beneficial than aggressive dimensionality reduction.

Future work could focus on advanced techniques for handling class imbalance, such as resampling methods or ensemble models, as well as exploring additional feature engineering strategies to further improve churn prediction performance.