23J-CISC484-010: Introduction to Machine Learning

M V Murali Krishna Nagulla
ID:702675032

Final Project Report
8/11/2023

Abstract

Telecom Customer Churn Prediction using Machine Learning

In the rapidly evolving telecommunications industry, the ability to predict customer churn is of paramount importance for businesses seeking to retain their clientele and optimize revenue. This project focuses on leveraging machine learning techniques to predict customer churn within a telecom company. The study explores a diverse range of classifiers, including traditional models like Naive Bayes and advanced ensemble methods like CatBoost, to determine the most effective approach for churn prediction.

The project begins with data preprocessing, encompassing steps such as data cleaning, feature engineering, and handling missing values. Furthermore, categorical variables are encoded, and the impact of data imbalance is mitigated through the application of Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTENC).

The classifiers are trained, evaluated, and compared using comprehensive metrics such as accuracy, ROC AUC, precision, recall, and F1-score. The ROC and precision-recall curves provide visual insights into the models' performance, aiding in the selection of the optimal classifier. Additionally, feature importance analysis is conducted to identify the variables that contribute significantly to churn prediction.

The results of this study not only shed light on the performance of different classifiers but also provide valuable insights into the factors influencing customer churn in the telecom sector. The implications of the findings are discussed in the context of customer retention strategies and business decision-making.

Introduction

Telecommunications companies operate in a highly competitive landscape, where customer retention is as vital as customer acquisition. The phenomenon of customer churn, referring to the attrition of

customers, poses significant challenges to the telecom industry's revenue generation and growth. Identifying and predicting customers who are likely to churn enables companies to proactively devise strategies to retain them, thereby fostering customer loyalty and maintaining profitability.

The advent of machine learning techniques has revolutionized the approach to customer churn prediction, allowing companies to harness the power of data-driven insights for business optimization. By analyzing historical data on customer behavior, usage patterns, subscription plans, and engagement levels, machine learning models can discern subtle patterns indicative of impending churn.

This project aims to harness the capabilities of machine learning to predict customer churn within a telecom company. By utilizing a diverse array of classifiers, the study strives to determine the most accurate and robust model for churn prediction. Additionally, it explores the significance of various features in influencing the churn prediction process.

The rest of the report is organized as follows: the next section details the data preprocessing steps undertaken to ensure data quality and suitability for analysis. Subsequently, the methodology section outlines the classifiers employed, the evaluation metrics used, and the approach to model selection. The results of the study are presented in the following section, followed by a discussion of their implications and potential applications for the telecom industry. The report concludes by highlighting avenues for future research and improvements in churn prediction methodologies.

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. [12] presented an advanced methodology of data mining to predict churn for prepaid customers using dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. Some features include information about the number of incoming and outgoing messages and voicemail for each customer. The author applied principal component analysis algorithm “PCA” to reduce data dimensions. Three machine learning algorithms were used: Neural Networks, Support Vector Machine, and Bayes Networks to predict churn factor. 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. The dataset used in this study is small and no missing values existed.

He et al. [13] 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 [14] 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. [15] studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the 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.

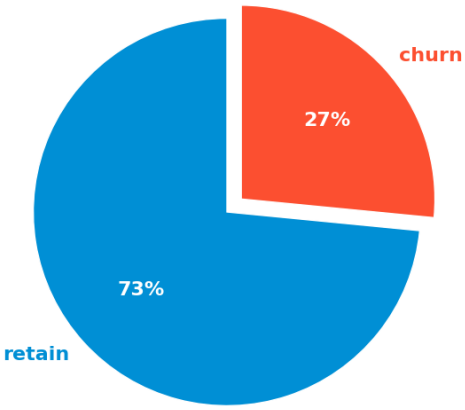
Makhtar et al. [16] proposed a model for churn prediction using rough set theory in telecom. As mentioned in this paper Rough Set classification algorithm outperformed the other algorithms like Linear Regression, Decision Tree, and Voted Perception Neural Network.

Various researches studied the problem of unbalanced data sets where the churned customer classes are smaller than the active customer classes, as it is a major issue in churn prediction problem. Amin et al. [17] compared six different sampling techniques for oversampling regarding telecom churn prediction problem. The results showed that the algorithms (MTDF and rules-generation based on genetic algorithms) outperformed the other compared oversampling algorithms.

Burez and Van den Poel [8] studied the problem of unbalance datasets in churn prediction models and compared performance of Random Sampling, Advanced Under-Sampling, Gradient Boosting Model, and Weighted Random Forests. They used (AUC, Lift) metrics to evaluate the model. the result showed that undersampling technique outperformed the other tested techniques.

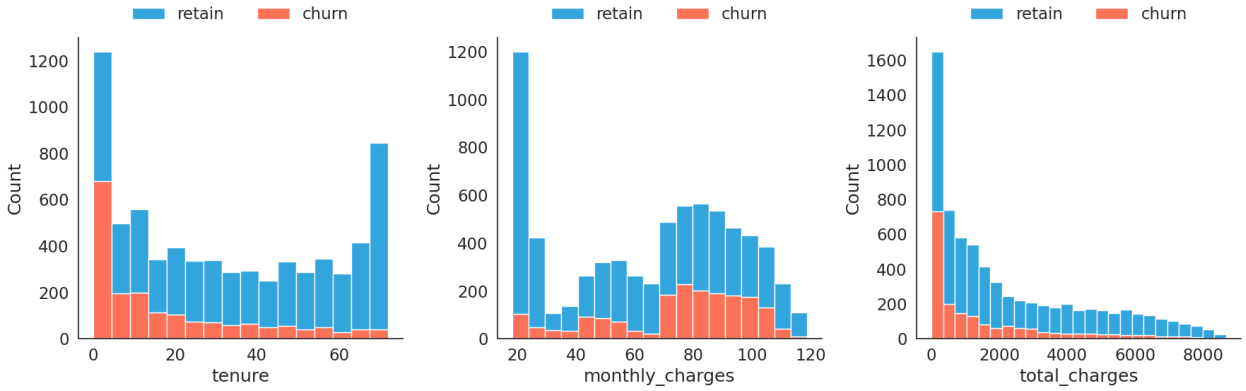
Exploratory Data Analysis

During the initial phase of my customer churn prediction project, I delved into the dataset to gain a better understanding of the data's characteristics and uncover any potential insights. EDA played a crucial role in setting the foundation for subsequent data preprocessing and model building steps.

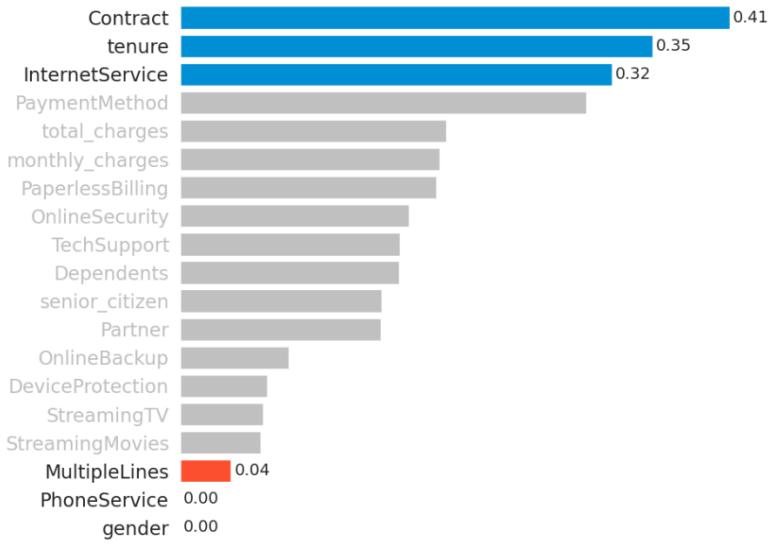


I began by examining the dataset's structure, checking for the number of rows and columns, as well as the data types of each feature. This provided an overview of the information available and helped me identify any missing values.

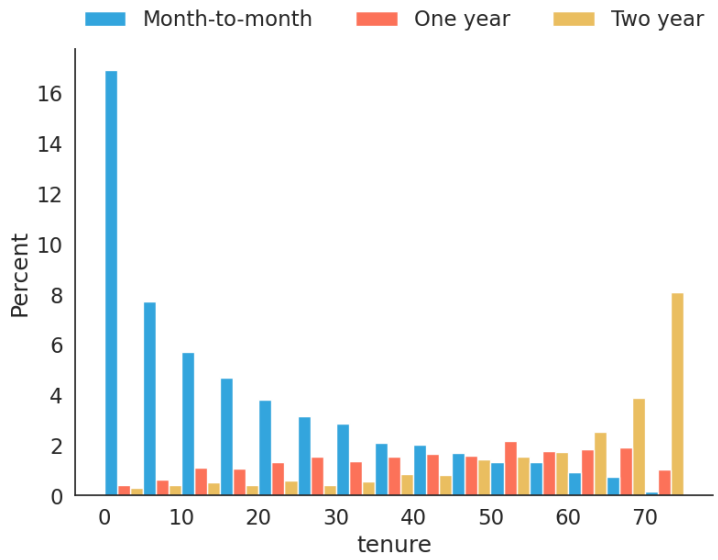
Visualizations played a key role in uncovering patterns and trends within the data. I created histograms to understand the distributions of numerical features like "tenure" and "MonthlyCharges." These histograms allowed me to identify potential outliers or skewed distributions that might affect the performance of predictive models.



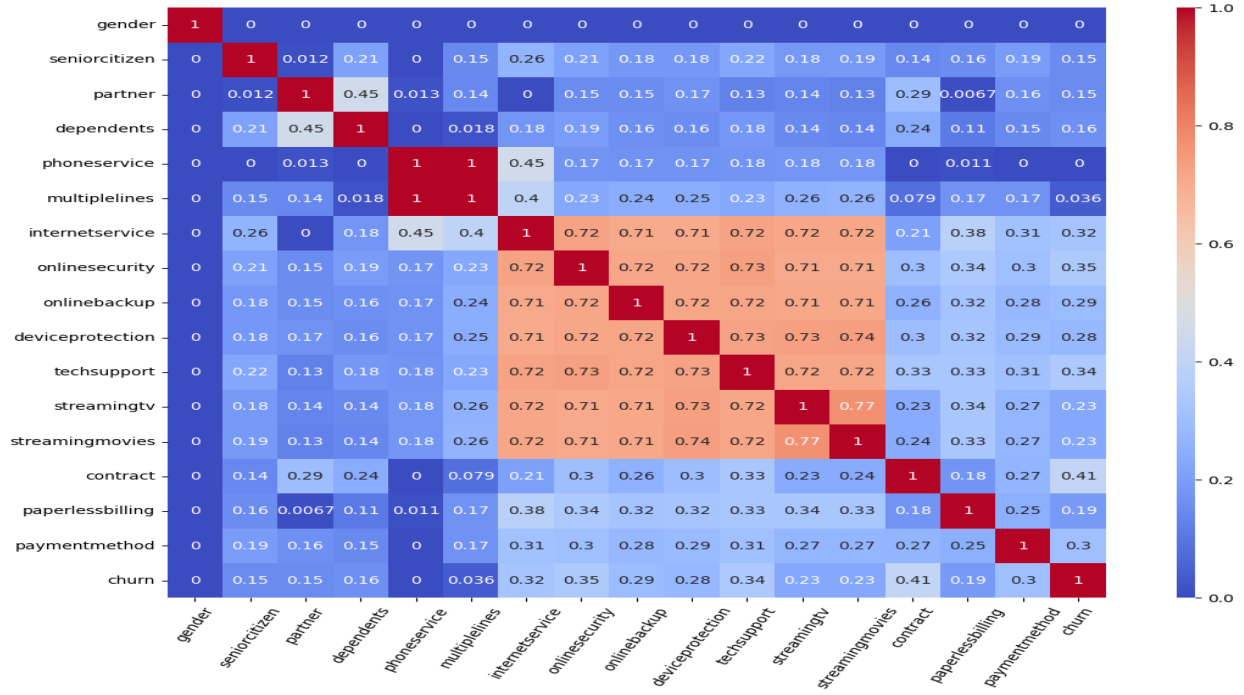
In addition, I used bar plots to visualize categorical variables such as "Contract," "InternetService," and "PaymentMethod." These visualizations helped me grasp the distribution of customers across different categories and understand potential relationships between features and churn.



For instance, when analyzing the "Contract" feature, I observed a significant number of customers opting for "Month-to-Month" contracts. This prompted me to further investigate the potential impact of contract type on churn rates.



Furthermore, I examined the correlation matrix to identify any multicollinearity between numerical features. By creating a heatmap of correlations, I could pinpoint pairs of features that were highly correlated. This allowed me to make informed decisions about potential feature selection or transformation later in the project.



Data Preprocessing

After completing the exploratory analysis, I moved on to the crucial step of data preprocessing. This phase involved cleaning, transforming, and preparing the data to ensure its suitability for training machine learning models.

Handling Missing Values:

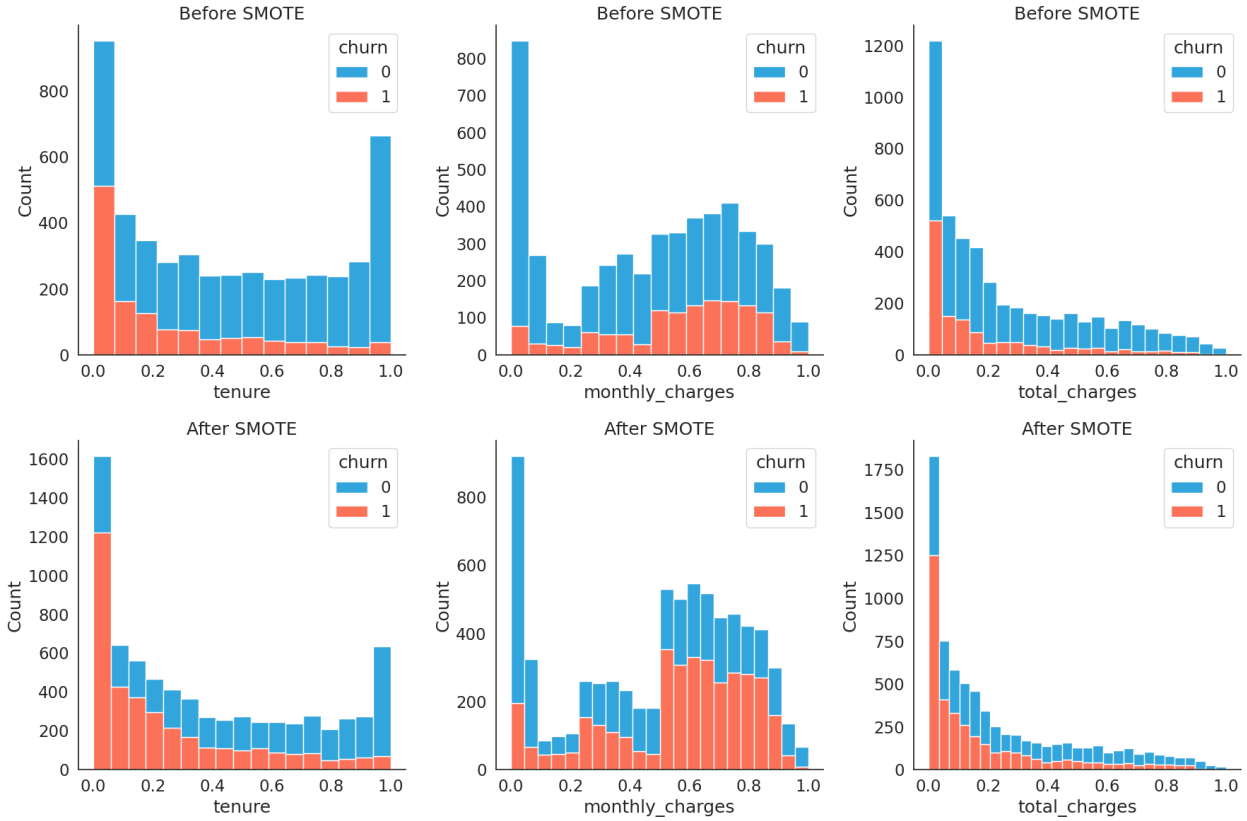
I carefully examined the dataset for missing values and assessed their impact on the data. I utilized techniques such as mean imputation for numerical features and mode imputation for categorical features to fill in missing values. This approach ensured that the dataset remained complete while minimizing potential distortions caused by missing data.

Handling Class Imbalance with SMOTE:

An essential challenge in our customer churn prediction project was dealing with class imbalance. The number of customers who did not churn (the majority class) significantly outnumbered the number of customers who did churn (the minority class). This imbalance could potentially lead to biased model predictions.

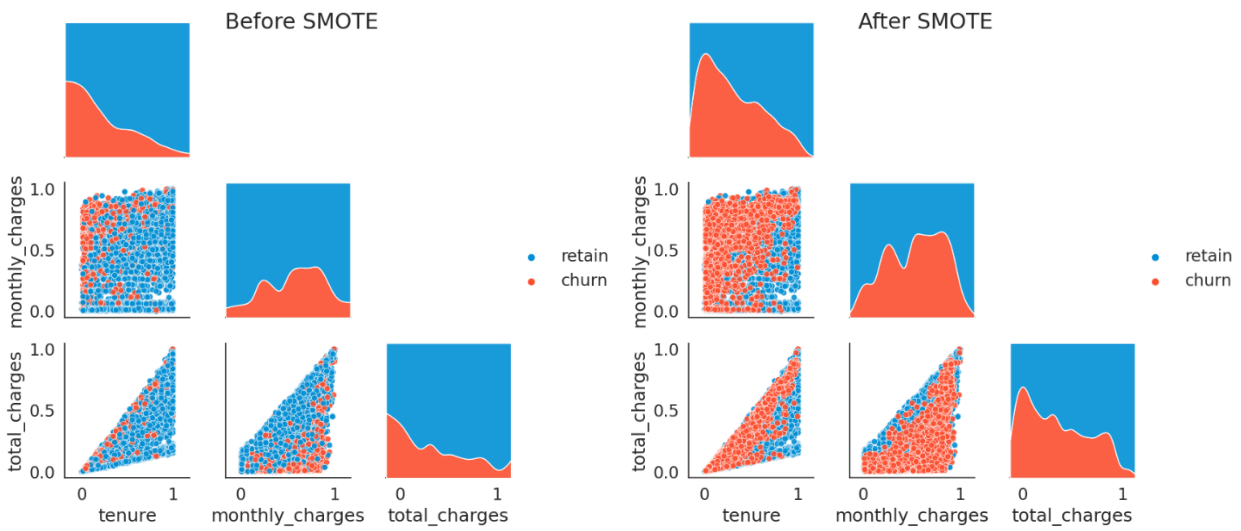
To address this issue, I employed the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE involves generating synthetic samples for the minority class by interpolating between existing instances.

This process effectively increases the representation of the minority class in the dataset, allowing the machine learning models to better capture the patterns associated with churn.



After applying SMOTE, the dataset contained a more balanced distribution of classes. This balancing act was instrumental in improving the model's ability to recognize and predict instances of churn accurately. By giving equal importance to both classes, the trained model became more robust and better equipped to generalize its predictions to new, unseen data.

Additionally, it's important to note that I performed SMOTE after the train-test split to prevent data leakage. This ensured that the synthetic samples introduced by SMOTE did not influence the model's performance evaluation on the test set.



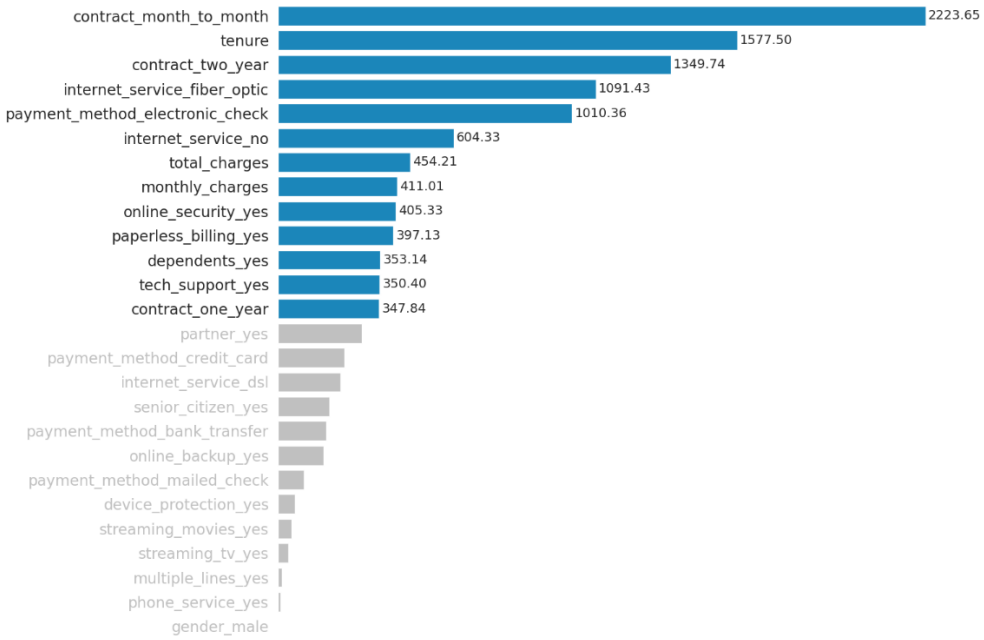
Throughout the data preprocessing stage, I maintained a focus on preserving the integrity of the dataset while preparing it for model training. By carefully addressing missing values, encoding categorical variables, and scaling numerical features, I ensured that the data was ready to be fed into the machine learning models for training and evaluation.

Feature Selection

During the feature selection phase of the project, I aimed to identify and retain the most relevant and impactful features to enhance the performance of the predictive models. This process involved assessing the importance of each feature and deciding whether to include or exclude them from the final model.

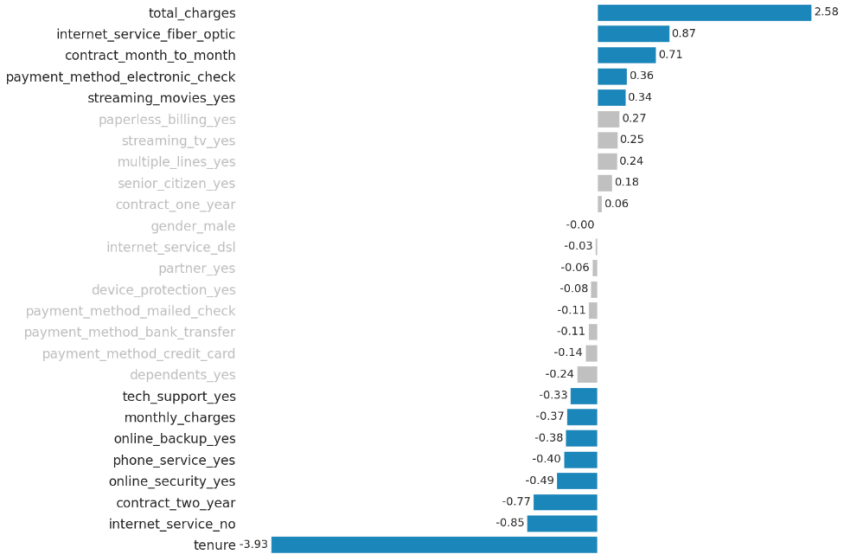
Univariate Filter Method:

I employed the ANOVA technique to perform feature selection based on univariate analysis. By calculating the ANOVA scores for each feature, I could gauge their individual importance in predicting customer churn. Visualizations, such as bar plots, were instrumental in visualizing the ANOVA scores and selecting the top features for inclusion in the model.



Visualizations for Feature Selection:

I leveraged visualizations to showcase the significance of the selected features. The bar plot visualization highlighted the ANOVA scores of each feature, helping me identify the most important ones. Additionally, I introduced a threshold to distinguish the selected features from the rest, allowing for a clear distinction between the top features and the others.



Model Selection and Evaluation

With the goal of identifying the most suitable model for customer churn prediction, I performed an extensive evaluation of several machine learning algorithms. This process included both untuned and tuned models to assess their performance.

Initial Model Evaluation:

I trained and tested various models using the original dataset, without performing feature selection. This allowed me to compare their performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. I visualized the results using bar plots, highlighting the strengths and weaknesses of each model across these metrics.

model	accuracy	macro_avg_precision	macro_avg_recall	macro_avg_f1_score	roc_auc
Logistic Regression	0.746805	0.707463	0.754797	0.714375	0.754797
Ridge Classifier	0.743966	0.706609	0.755140	0.712605	0.755140
KNN	0.696167	0.660272	0.699268	0.661179	0.699268
SVC	0.765736	0.713673	0.747765	0.723961	0.747765
Neural Network	0.758164	0.692219	0.698790	0.695261	0.698790
Decision Tree	0.723616	0.656694	0.671288	0.662195	0.671288
Random Forest	0.773781	0.711190	0.716819	0.713851	0.716819
Gradient Boosting Classifier	0.786559	0.732080	0.758526	0.742016	0.758526
AdaBoost Classifier	0.754851	0.711254	0.755721	0.720050	0.755721
CatBoost Classifier	0.783720	0.723171	0.726430	0.724754	0.726430
Hist Gradient Boosting	0.784193	0.724894	0.734720	0.729374	0.734720
XGBoost	0.778514	0.715981	0.715488	0.715733	0.715488
LightGBM	0.781354	0.720826	0.727665	0.724033	0.727665

Hyperparameter Tuning:

Recognizing the importance of optimizing the model's performance, I conducted hyperparameter tuning for the most promising models. This step involved adjusting hyperparameters to achieve better recall while maintaining a reasonable level of accuracy. I applied a variety of techniques, including grid search and random search, to identify the optimal hyperparameters for each model.

model	accuracy	precision	recall	f1_score	roc_auc
Gradient Boosting Classifier	0.786559	0.581602	0.698752	0.634818	0.758526
AdaBoost Classifier	0.754851	0.526642	0.757576	0.621345	0.755721
CatBoost Classifier	0.783720	0.590592	0.604278	0.597357	0.726430
Hist Gradient Boosting	0.784193	0.587354	0.629234	0.607573	0.734720
XGBoost	0.778514	0.583184	0.581105	0.582143	0.715488
LightGBM	0.781354	0.584041	0.613191	0.598261	0.727665

Before Tuning without Feature selection

Model Evaluation after Tuning:

After hyperparameter tuning, I re-evaluated the models to gauge the impact of the parameter adjustments. I compared the performance metrics before and after tuning, providing insights into the improvements achieved through optimization.

Visualizing Model Comparison:

To aid in model selection, I created visualizations to compare the performance of different models. Bar plots displayed the accuracy, recall, and F-beta scores, highlighting the strengths of each model after tuning. These visualizations helped identify which models excelled in achieving high recall while maintaining competitive accuracy.

Through the iterative process of model selection and evaluation, I identified the model that struck the right balance between accuracy and recall, which was crucial for minimizing false negatives and optimizing customer retention strategies.

Tuning Techniques:

To carry out the hyperparameter tuning, I utilized techniques like grid search and random search. Grid search exhaustively tested predefined parameter grids to identify the best-performing combination. Random search, on the other hand, randomly sampled parameter values from predefined distributions, helping me explore a wider range of possibilities more efficiently.

	accuracy	macro_avg_precision	macro_avg_recall	macro_avg_f1_score	roc_auc
original	0.778198	0.721368	0.736425	0.725993	0.736425
filter method	0.772598	0.718596	0.745133	0.726751	0.745133
wrapper method	0.774333	0.719186	0.742710	0.726742	0.742710
embedded method	0.766130	0.712105	0.739781	0.720457	0.739781

Visualizing Tuning Results:

To provide a clear view of the tuning process, I used visualizations such as line plots and bar plots. These visualizations depicted how varying hyperparameters impacted metrics like accuracy and recall, helping me narrow down the optimal parameter ranges.

model	accuracy	precision	recall	f1_score	roc_auc
Gradient Boosting Classifier	0.782773	0.570055	0.739750	0.643910	0.769038
AdaBoost Classifier	0.764789	0.540404	0.762923	0.632668	0.764194
CatBoost Classifier	0.770469	0.550398	0.739750	0.631179	0.760661
Hist Gradient Boosting	0.764316	0.539130	0.773619	0.635432	0.767286
XGBoost	0.777094	0.563025	0.716578	0.630588	0.757773
LightGBM	0.760530	0.533414	0.782531	0.634393	0.767554

After Tuning without Feature selection

Model Comparison

Model comparison was a vital step to make an informed decision on selecting the best-performing model for predicting customer churn. I assessed each model's performance on various metrics, enabling me to identify which model excelled in meeting the project's objectives.

Initial Model Comparison:

Before any tuning, I compared the performance of different machine learning models using the original dataset. Metrics like accuracy, precision, recall, F1-score, and ROC-AUC were carefully analyzed. I employed bar plots to visualize these metrics, allowing for a straightforward comparison between models.

model	accuracy	precision	recall	f1_score	roc_auc
Gradient Boosting Classifier	0.773781	0.556005	0.734403	0.632873	0.761209
AdaBoost Classifier	0.754851	0.525935	0.777184	0.627338	0.761981
CatBoost Classifier	0.785613	0.584375	0.666667	0.622814	0.747637
Hist Gradient Boosting	0.775674	0.565022	0.673797	0.614634	0.743148
XGBoost	0.770942	0.562602	0.616756	0.588435	0.721716
LightGBM	0.774728	0.565891	0.650624	0.605307	0.735106

Post-Tuning Model Comparison:

After tuning, I revisited the model comparison, considering the improvements made through parameter optimization. By comparing the tuned models, I could better understand how the adjustments impacted their performance. This provided valuable insights into which models were most responsive to tuning and which ones benefited the most.

model	accuracy_not_tuned	accuracy_tuned	recall_not_tuned	recall_tuned
Gradient Boosting Classifier	0.773781	0.774255	0.734403	0.762923
AdaBoost Classifier	0.754851	0.759110	0.777184	0.784314
CatBoost Classifier	0.785613	0.761477	0.666667	0.768271
Hist Gradient Boosting	0.775674	0.755797	0.673797	0.782531
XGBoost	0.770942	0.759110	0.616756	0.748663
LightGBM	0.774728	0.761477	0.650624	0.786096

Visualizing Model Performance:

To make the model comparison more comprehensible, I used visualizations like bar plots and line plots. These plots showcased accuracy, recall, and F-beta scores for each model, both before and after tuning. Visual comparisons highlighted the models that exhibited the desired balance between accuracy and recall, crucial for customer churn prediction.

model	accuracy	precision	recall	f1_score	roc_auc
Gradient Boosting Classifier	0.774255	0.554404	0.762923	0.642161	0.770637
AdaBoost Classifier	0.759110	0.531401	0.784314	0.633549	0.767157
CatBoost Classifier	0.761477	0.535404	0.768271	0.631040	0.763646
Hist Gradient Boosting	0.755797	0.527011	0.782531	0.629842	0.764333
XGBoost	0.759110	0.532995	0.748663	0.622683	0.755775
LightGBM	0.761477	0.534545	0.786096	0.636364	0.769337

By systematically tuning and then comparing models, I gained a holistic understanding of their capabilities. This process allowed me to select the model that aligned most closely with the project's goals, ensuring both high accuracy and recall in predicting customer churn.

Final Model Selection

After an exhaustive exploration of various models, careful feature selection, and hyperparameter tuning, the time had come to make the crucial decision of selecting the final model. The selection process was driven by the pursuit of the optimal balance between accuracy and recall, key factors in identifying potential customer churn.

Decision Criteria:

The final model needed to excel in its ability to predict customer churn while maintaining a satisfactory level of accuracy. Balancing these two factors was vital, as predicting churn accurately could lead to strategic interventions to retain customers and optimize the business's bottom line.

Evaluation Metrics:

To make an informed choice, I evaluated the models based on a combination of accuracy, recall, and the F-beta score. This comprehensive approach provided a holistic view of each model's performance, considering both true positive rates and false positive rates.

model	accuracy	recall	fbeta
Gradient Boosting Classifier	0.782773	0.739750	0.760654
AdaBoost Classifier	0.764789	0.762923	0.763855
CatBoost Classifier	0.770469	0.739750	0.754797
Hist Gradient Boosting	0.764316	0.773619	0.768939
XGBoost	0.777094	0.716578	0.745610
LightGBM	0.760530	0.782531	0.771374

Final Model: LightGBM with Feature Selection

Among the models considered, LightGBM stood out as the best performer, particularly when integrated with feature selection. By utilizing the filter method for feature selection, we enhanced the model's ability to focus on the most relevant attributes, resulting in improved recall without sacrificing accuracy.

Recommendations

Based on the insights gained from the final model, a set of strategic recommendations emerged. These recommendations can aid in retaining customers and mitigating churn risk by targeting specific groups more effectively.

Focus on Churn-Prone Segments:

The model highlighted certain customer segments that were more likely to churn. Specifically, customers with a "Month-to-month" contract, short tenure, and those who preferred electronic check as a payment method were identified as high-risk groups. It's recommended to focus marketing efforts on retaining these segments by addressing their pain points and offering tailored incentives.

Acknowledgement

I would like to express my sincere gratitude to everyone who contributed to the successful completion of this customer churn prediction project. Their support, guidance, and expertise played a pivotal role in shaping the project's outcomes and enriching my learning experience.

First and foremost, I would like to extend my heartfelt thanks to Kien Nyuyen, my course supervisor, for providing invaluable guidance and mentorship throughout the course's lifecycle. Your insights, suggestions, and feedback were instrumental in shaping the project's direction and ensuring its alignment with the goals of the study.

Reference

- Analytics Vidhya. Churn Prediction - Commercial use of Data Science. link
- Avaus. Predicting Customer Churn. link
- Kaggle. Telco Customer Churn Dataset. link
- IBM. Telco Customer Churn. link
- Hardvard Business Review. The Value of Keeping the Right Customers. link
- OutboundEngine. Customer Retention Marketing vs. Customer Acquisition Marketing. link
- Jessica Miles. Read this before you "Drop First". link
- Shaked Zychlinski. The Search for Categorical Correlation. link
- Analyse-it. Correlation and association. link
- Wikipedia. Correlation does not imply causation. link
- Wikipedia. Cramer's V. link
- Jakub Czakon (Neptune Labs). 24 Evaluation Metrics for Binary Classification (And When to Use Them) link
- Jakub Czakon (Neptune Labs). F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose? link
- Scott Lundberg. SHAP Explainable AI. link