Churn Prediction Model

By:

Diego Rueda

Professor:

Ashish

Course:

Applied Artificial Intelligence and Machine Learning

Business Analytics 2024

St Lawrence College.

Kingston, On

Canada

Introduction

In the highly competitive telecom industry, customer retention is crucial for growth and profitability. Customers churn, the decision from a customer to stop using a company's services, represent a significant challenge for corporations and being able to predict it opens the opportunity for strategies for retention. Understanding and predicting churn can enable telecom companies to proactively address issues, enhance customer satisfaction, and implement targeted strategies for retention.

This report focuses on predicting customer churn using advanced Artificial Intelligence (AI) algorithms implemented in Python. By leveraging data-

driven insights, the aim is to identify patterns and indicators that predict customer departure (Churn). This predictive approach will help in creating effective strategies for customers retention and improving business performance for TeleCom Inc.

The study will utilize a dataset from TeleCom Inc. Canada, encompassing a variety of customer attributes and service usage metrics. The process will include data preprocessing, feature engineering, model selection, and evaluation. The goal is to predict customer churn.

The data

The data provided contain the following information:

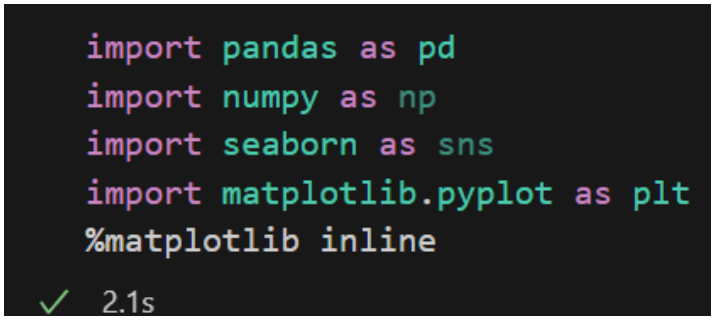
1. CustomerID: A unique identifier assigned to each customer.
2. Gender: The gender of the customer.
3. SeniorCitizen: Indicates if the customer is a senior citizen.
4. Partner: Indicates if the customer has a partner.
5. Dependents: Indicates if the customer has dependents.
6. Tenure: The number of months the customer has been with the company.
7. PhoneService: Indicates if the customer has phone service.
8. MultipleLines: Indicates if the customer has multiple phone lines.
9. InternetService: Indicates if the customer has Internet Service.
10. OnlineSecurity: Indicates if the customer has online security service.
11. OnlineBackup: Indicates if the customer has online backup service.
12. DeviceProtection: Indicates if the customer has device protection service.
13. TechSupport: Indicates if the customer has tech support service.
14. StreamingTV: Indicates if the customer has streaming TV service.
15. StreamingMovies: Indicates if the customer has streaming movies service.
16. Contract: The type of contract the customer has.
17. PaperlessBilling: Indicates if the customer has opted for paperless billing.
18. PaymentMethod: The method the customer uses for payment.
19. MonthlyCharges: The amount charged to the customer monthly.
20. TotalCharges: The total amount charged to the customer.
21. Churn: Indicates if the customer has churned.

Feature engineered column:

22. TotalCharges_Intensity: The amount customer gave the company over the time as customers.

Python

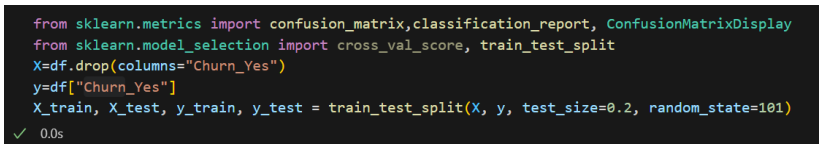
Using Python, we first load the data and perform initial checks and cleaning process. For this purpose, we utilize several python libraries like Pandas, Numpy, Seaborn and Matplotlib for data manipulation, numerical operations, for data visualization, and creating detailed plots. These tools collectively ensure our dataset is accurate, consistent, and ready for subsequent analysis and training of the AI models.



With the data loaded we can proceed to do some cleaning, and check for null values. Additionally, we transform the categorical elements using dummies and scale our numerical data points using the function StandardScaler from the Sklearn Library.

Prediction

Our primary objective is to predict whether a customer will leave the organization, thus churn is the selection as our prediction target. To achieve this, we designate 'Churn' as our dependent variable and split the dataset into training and testing subsets to facilitate model training and evaluation. Using the Scikit-learn (SKlearn) library, we ensure that the data is partitioned correctly.



Random Forest Classifier

First, we utilize the Random Forest Classifier for our initial model. This ensemble learning method constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. By leveraging the Random Forest Classifier, we can improve predictive accuracy and control over-fitting. Using the powerful machine learning library, Sklearn (Scikit-learn), we implement this model to capitalize on its robustness and ability to handle large datasets with higher dimensional spaces effectively.

```
from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier()
rf_model.fit(X_train, y_train)
rf_predictions = rf_model.predict(X_test)
```
✓ 0.9s

Logistic Regression

Second, we employ Logistic Regression, a well-established algorithm that performs well in binary classification tasks. Logistic Regression estimates the probability of a binary response based on one or more predictor variables. Implemented using Sklearn, providing a straightforward interpretation of the coefficients and facilitating the prediction of customer churn with a high degree of accuracy.

```
from sklearn.linear_model import LogisticRegression
logistic = LogisticRegression().fit(X_train, y_train)
predict = logistic.predict(X_test)
```
✓ 0.0s

SVC

Third, we utilize the Support Vector Classifier (SVC), an advanced algorithm designed for classification problems. SVC works by finding the hyperplane that best divides a dataset into classes, maximizing the margin between different classes. This method is particularly useful for both linear and non-linear data classifications. With the implementation through Sklearn, ensuring robust and precise churn prediction.

```
from sklearn.svm import SVC
model = SVC()
model.fit(X_train,y_train)
predictions = model.predict(X_test)
```
✓ 1.0s

Neural Network

Fourth, we develop a Neural Network model, leveraging the capabilities of TensorFlow and Keras. Neural Networks are a class of machine learning algorithms inspired by the structure and function of the human brain, perfect for handling complex patterns and relationships in data. Utilizing TensorFlow and Keras, we can efficiently build, train, and evaluate our model. This approach allows us to capture intricate data dependencies and interactions, enhancing our possibilities to predict customer churn with higher accuracy and generalization to unseen data

```
model=Sequential()

model.add(Dense(units=40,activation="relu"))#hiden
model.add(Dense(units=20,activation="relu"))#hiden
model.add(Dense(units=40,activation="tanh"))#hiden

model.add(Dense(units=1 ,activation="sigmoid")) #output
model.compile(optimizer="adam",loss="binary_crossentropy")
```
✓ 0.0s

For this model, we create a network of 3 hidden layers. To ensure that our model don't over fit. We create a stopping variable.

```
early_stop  = EarlyStopping(monitor  = "val_loss",
                            patience = 40,
                            verbose  = 1,
                            mode="min")
```

✓ 0.0s

```
model.fit(X_train,
          y_train,
          epochs=500,
          validation_data = (X_test,y_test),
          callbacks = early_stop)
```

✓ 13.1s

Result

We compare the four models together and evaluate their performances in predicting customer churn. It is that all four models encounter challenges in predicting churn. Despite their individual strengths, the Random Forest Classifier, Logistic Regression, Support Vector Classifier (SVC), and Neural Network models each exhibit certain limitations when applied to our dataset. This can be caused by as imbalanced classes, the complexity of customer behavior, and the inherent noise in the data contribute to these difficulties. Given the need to correctly predict churn we decide that the best model is Logistic Regression, with the highest f score.

Random Forest Classifier

	precision	recall	f1-score	support
Will not Churn	0.84	0.89	0.87	1052
Will Churn	0.62	0.50	0.55	355
accuracy			0.80	1407
macro avg	0.73	0.70	0.71	1407
weighted avg	0.78	0.80	0.79	1407

Logistic Regression

	precision	recall	f1-score	support
Will not Churn	0.86	0.90	0.88	1052
Will Churn	0.66	0.55	0.60	355
accuracy			0.81	1407
macro avg	0.76	0.73	0.74	1407
weighted avg	0.81	0.81	0.81	1407

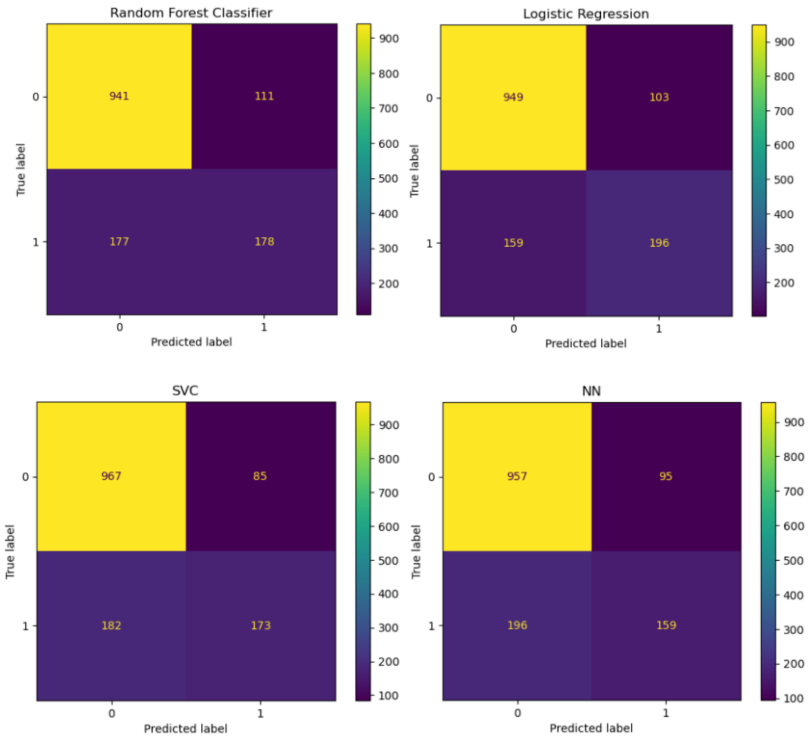
SVC

	precision	recall	f1-score	support
Will not Churn	0.84	0.92	0.88	1052
Will Churn	0.67	0.49	0.56	355
accuracy			0.81	1407
macro avg	0.76	0.70	0.72	1407
weighted avg	0.80	0.81	0.80	1407

NN

	precision	recall	f1-score	support
Will not Churn	0.83	0.91	0.87	1052
Will Churn	0.63	0.45	0.52	355
accuracy			0.79	1407
macro avg	0.73	0.68	0.70	1407
weighted avg	0.78	0.79	0.78	1407

When we compare the results using the confusion matrix, it provides us with an interesting perspective on the models' performance. The confusion matrix not only illustrates the number of true positives, true negatives, false positives, and false negatives for each model but also highlights where each model excels and where they falter. This detailed breakdown allows us to pinpoint specific areas of misclassification and understand the balance between sensitivity and specificity. Such insights are crucial for refining our models and developing targeted strategies to improve their predictive accuracy in future iterations.



Conclusion

In conclusion, our comprehensive analysis utilizing four distinct models Random Forest Classifier, Logistic Regression, Support Vector Classifier, and Neural Networks, provides valuable insights into the complexities of customer churn prediction. While each model demonstrates unique strengths and specific challenges, the evaluation through confusion matrices offers insights into understanding their results. Despite that predicting churn is a difficult due to many factors, these models lay a foundation for further refinement. Future efforts should focus on addressing data imbalances, enhancing feature engineering, and incorporating additional data sources to improve model accuracy. Finaly, this study highlights the importance of continuous model evaluation and adaptation in the ever-evolving landscape of customer behavior analytics for TeleCom Inc. This is interesting for today's telecommunication competitions landscape.