Documentation Report: Customer Churn Prediction

Overview

This paper describes the creation, assessment, and implementation of a Support Vector Classifier (SVC) model for forecasting customer churn that was selected from the phase 2 modeling process (candidate models are given in the image below). To improve the model's performance, the emphasis was on hyperparameter optimization and regularization.

Model	Precision	Recall	F1-Score	Support	Accuracy	ROC-AUC	ROC-AUC (Class 0)	ROC-AUC (Class 1)	Average Precision
Baseline Logistic Reg	0.85	0.90	0.87	1031	0.81		0.86	0.86	0.69
Optimized Logistic Reg	0.85	0.90	0.87	1031	0.81		0.86	0.86	0.70
Baseline SVC	0.84	0.92	0.88	1031	0.81	0.81			0.66
Optimized SVC	0.84	0.92	0.88	1031	0.81	0.81			0.66
Baseline Random Forest	0.84	0.92	0.88	1031	0.81		0.84	0.84	0.67
Optimized Random Forest	0.84	0.92	0.88	1031	0.81		0.84	0.84	0.67
Baseline Bagging Clf	0.83	0.91	0.87	1031	0.79		0.81	0.81	0.61
Optimized Bagging Clf	0.84	0.90	0.87	1031	0.80		0.84	0.84	0.66
Baseline Naive Bayes	0.90	0.72	0.80	1031	0.74		0.84	0.84	0.66
Optimized Naive Bayes	0.90	0.72	0.80	1031	0.74		0.84	0.84	0.66
Neural Network	0.82	0.94	0.88	1031	0.80	0.69			0.46
Optimized Neural Net	0.84	0.92	0.88	1031	0.81	0.72			0.49

Hyperparameter Tuning

Methodology

- **Approach**: Grid search was used to tune the SVC model's hyperparameters.
- **Parameters Considered**:
  - **C (Regularization Parameter)**: This parameter controls the trade-off between increasing the margin size and ensuring that the samples are accurately categorized.
  - **Kernel**: Modifies the feature space. 'Linear,' 'rbf,' 'poly,' and'sigmoid' were among the options considered.
- **Validation Method**: The performance of several hyperparameter combinations was evaluated using 5-fold cross-validation.

Results

- **Optimal Hyperparameters**:
  - **C**: 1
  - **Kernel**: 'rbf'

Model Training

- **Configuration**: The SVC model was trained with `C=1` and `Kernel='rbf'`.
- **Data Used**: Training was conducted using the provided training dataset (`X_train`, `y_train`).

Model Effectiveness

Performance Metrics

- **Accuracy**: Achieved an accuracy of 81% (0.81), indicating the proportion of correct predictions.
- **Precision and Recall**:
  - **Class 0**:
    - Precision: 0.84 (Model is correct 84% of the time when predicting class 0).
    - Recall: 0.92 (High ability to recognize true positives for class 0).
  - **Class 1**:
    - Precision: 0.69 (Lower reliability for class 1 predictions).
    - Recall: 0.52 (Less effective in recognizing true positives for class 1).
- **F1-Score**:
  - Class 0: 0.88
  - Class 1: 0.60
- **ROC-AUC**: The model scored 0.81, indicating good discriminatory power between classes.
- **Average Precision**: 0.66, reflecting the precision-recall trade-off.

Insights

- The SVC model's performance was greatly enhanced by hyperparameter adjustment, notably with the RBF kernel and a regularization value of C=1.
- There was a significant performance difference between class 0 and class 1 forecasts, with class 0 predictions being more accurate.
- The model has significant predictive capability, however it has to be refined further for class 1 predictions.

Recommendations

- **Model Improvement**:
  - Look at other kernels or regularization settings to improve class 1 predictions.
  - To equalize the model's performance across classes, consider adding features or using data preparation techniques.
- **Further Analysis**:
  - Examine the properties of class 1 to see why the model is less successful for these predictions.
- **Model valuation**:
  - Additional measures, such as a confusion matrix, might give further insights into the model's sorts of mistakes.

- **Deployment Considerations**:
    - Before deploying the model, make sure it works effectively across several client segments and operational settings.

Streamlit Web Application Functionality

Overview

The Streamlit online application is an interactive interface that allows end users to use the SVC model to anticipate customer attrition. It provides two key functions:

1. **Single Customer Prediction**
2. **Batch Prediction via CSV Upload**

Single Customer Prediction

Functionality

- Users can enter information about a particular consumer into the online application.
- The input fields correspond to the SVC model's characteristics.
- After entering the details, the user can submit the data for prediction.

Process Flow

1. **Input Data**: The user enters the customer's information into the supplied input forms.
2. **Submit for Prediction**: The SVC model processes the data after it is submitted.
3. **Display Results**: The churn prediction result is displayed in the application, showing whether the client is likely to churn or not.

Batch Prediction via CSV Upload

Functionality

- Users can submit a CSV file providing information for multiple clients.
- The CSV file should have the format and structure that the model expects, with columns matching to client characteristics.

Process Flow

1. **CSV Upload**: The user uploads a CSV file with the customer data batch.
2. **Data Processing**: The SVC model is used by the program to process the full batch.
3. **Visualization and Results**:
    - A pie chart visualization is created, displaying the percentage of customers who are likely to churn vs those who are not.
    - Alongside the pie chart, a detailed forecast result for each client is provided, classifying them as 'Churn: Yes' or 'Churn: No'.

User Interface

- The Streamlit web application is intended to be simple and easy to use.
- Each input field and feature has clear instructions and labeling.
- To assist users in the event of improper data entry or file format difficulties, error handling and user feedback systems are implemented.

Technical Considerations

- **Data Handling**: To correspond with the model's needs, proper validation and preprocessing are done to user inputs.
- **Security and Privacy**: The program protects user data privacy and security by not storing personal or sensitive information.
- **Performance**: Optimized for efficient processing, including batch predictions, in order to offer rapid and accurate results.

Reccomendations based on problem statement

- **Targeted Intervention Strategies:** Users may rely on the model to identify clients at high risk of churn with an accuracy of 81%. Businesses can create tailored intervention methods for these consumers, such as specialized offers, better customer service, or addressing particular problems that may be causing churn.
- **Resource Optimization:** Companies may spend resources more efficiently if they can precisely forecast which consumers are likely to churn. Rather of distributing resources broadly across all clients, companies may concentrate on those indicated by the model, resulting in more cost-effective retention methods.
- **Customer Segmentation for Tailored Communication:** The accuracy and recall values for each class indicate that the model may be utilized for customer segmentation. Based on the model's categorization, businesses may modify their communication and retention tactics, providing more targeted and relevant interactions.
- **Performance Improvement for Class 1 Predictions:** Because the model has poorer precision and recall for class 1, more study and refining are required. Users might concentrate on collecting additional data on churning consumers or using other modeling strategies to enhance these KPIs.
- **Incorporating Real-Time Data:** Extending the research to include real-time data processing might improve the forecasting capabilities of the model. Businesses would be able to respond more immediately if a consumer showed indicators of future turnover.
- **Integration with CRM Systems:** By integrating this model with current Customer Relationship Management (CRM) systems, the churn prediction process may be automated, making it a seamless component of the customer management workflow.
- **Exploring Additional Variables:** Investigating other churn-influencing factors or attributes, such as customer satisfaction scores, service usage patterns, or competitor offers, might improve the model's accuracy and give deeper insights.

- **Predictive Maintenance:** The model might be used to anticipate not only customer turnover but also regions where service outages may lead to greater churn, enabling for predictive maintenance in industries such as telecommunications or software services.
- **Developing Customized Retention Offers:** Companies might construct personalized retention incentives or loyalty programs based on the model's findings to engage consumers who are expected to churn.
- **Longitudinal Analysis:** Longitudinal studies to track the efficacy of various retention tactics over time, as indicated by the model, might give significant input for further refining the model and the strategies based on it.

Conclusion

The improved SVC model predicts customer attrition with good outcomes. However, the observed disparity in performance between classes demands more modifications. Continuous model review and adaptation are critical for sustaining relevance and accuracy in real-world circumstances.

Using the power of the optimized SVC model, the Streamlit online application offers as an accessible and efficient solution for both individual and batch customer churn forecasts. This program is intended to meet a wide range of user requirements by providing simplicity in single predictions and efficiency in handling large amounts of data, all accompanied by meaningful visuals. It is a great tool for companies trying to leverage machine learning for client retention initiatives.