CS571 Comprehensive Customer Churn Analysis: From Exploration to Prediction

Kulkarni, Sanket
Illinois Institute of Technology
Chicago, USA
skulkarni27@hawk.iit.edu or A20537896

Hujare, Ameya
Illinois Institute of Technology
Chicago, USA
ahujare@hawk.iit.edu or A20545367

Pawar, Deep
Illinois Institute of Technology
Chicago, USA
dpawar3@hawk.iit.edu or A20545137

Reddy, Srinath
Illinois Institute of Technology
Chicago, USA
Msrinathreddy@hawk.iit.edu or A20561409

***Abstract*—Customer churn is a significant challenge for subscription-based industries, particularly in telecommunications, where retaining customers is crucial for profitability. This study investigates churn behavior using the Telco Customer Churn dataset to identify factors influencing customer attrition and develop predictive solutions. A structured methodology was adopted, including data preprocessing, exploratory analysis, dimensionality reduction using PCA and t-SNE, and clustering techniques such as K-Means and DBSCAN to segment customers. Predictive models, including Random Forest and Logistic Regression, were trained and optimized with hyperparameter tuning and cross-validation, addressing challenges like class imbalance. The analysis revealed key patterns, such as the impact of tenure, contract type, and payment methods on churn likelihood, offering actionable insights for retention strategies. The results demonstrate the potential of data-driven approaches to enhance customer retention, laying a foundation for real-time prediction and personalized interventions in future applications.**

Keywords: Customer churn, Predictive modeling, PCA, t-SNE, K-Means clustering, DBSCAN, Random Forest, Logistic Regression, Customer churn, Telecommunications

I. INTRODUCTION

Customer churn is a critical challenge faced by businesses, especially in subscription-based industries like telecommunications. Understanding why customers leave and identifying patterns in their behavior can provide valuable insights for improving customer retention and maximizing profitability. This project focuses on analyzing the Telco Customer Churn dataset to explore factors contributing to customer churn and develop predictive models to mitigate it.

The primary objective of this project is to leverage data-driven techniques to gain insights into the relationships between customer features and churn behavior. By employing a combination of exploratory data analysis, visualization strategies, unsupervised learning, and predictive modeling, this project aims to achieve the following:

- Understand the Dataset: Investigate and preprocess the dataset to identify correlations, dependencies, and the overall structure of features and their impact on the churn target.
- Visualize Patterns: Utilize advanced dimensionality reduction techniques like PCA, UMAP, and t-SNE to uncover patterns in high-dimensional data.
- Model Churn: Build and evaluate machine learning models to predict churn, applying rigorous cross-validation to ensure robust performance.
- Improve Model Performance: Explore potential strategies to enhance prediction accuracy, such as feature engineering, regularization, and experimenting with complex models.

Through this analysis, the project not only aims to deliver actionable insights into customer behavior but also proposes practical strategies for reducing churn rates, which can be instrumental for business decision-making in the telecommunications sector.

II. OBJECTIVES

The primary objectives of this project are:

1) **Understand Customer Churn Behavior:**
   - Investigate key factors influencing customer churn in the telecommunications industry.
   - Analyze relationships between customer attributes and their likelihood of churn.
2) **Perform Data Exploration and Visualization:**
   - Utilize various visualization techniques to uncover patterns and relationships in the data.
   - Apply dimensionality reduction methods such as PCA, UMAP, and t-SNE to explore data structure and gain deeper insights.
3) **Apply Unsupervised Learning Techniques:**

- Cluster customers based on their features to identify similar groups and explore patterns independent of the target variable.

4) **Develop Predictive Models:**
   - Train baseline machine learning models to predict churn effectively.
   - Evaluate model performance using appropriate metrics and cross-validation techniques.
5) **Optimize Model Performance:**
   - Identify strategies for improving model accuracy, such as feature selection, regularization, and hyperparameter tuning.
   - Experiment with advanced models to explore potential performance improvements.
6) **Gain Business Insights:**
   - Translate data-driven findings into actionable recommendations for reducing churn and improving customer retention.
   - Propose interventions that can be implemented by the organization to address key factors contributing to churn.
7) **Foster Continuous Improvement:**
   - Conduct additional experiments to test hypotheses and validate insights.
   - Provide a framework for future analysis and model enhancements.

III. DATABASE DESCRIPTION

The **Telco Customer Churn** dataset is a publicly available dataset often used for customer behavior analysis in the telecommunications sector. It provides detailed information about customers, their services, account details, and whether they have churned (left the company) or not. Below is a detailed description of the database:

1) **Overview**
   - **Total Records:** The dataset contains approximately **7,043 records**, each representing a unique customer.
   - **Features:** The dataset has **21 columns**, comprising numerical, categorical, and target variables.
   - **Target Variable:** The column Churn indicates whether a customer has left the company (Yes) or not (No).
2) **Key Feature Categories** The dataset includes features broadly grouped into the following categories:
   a) **Customer Demographics:**
      - **Gender:** Male or Female.
      - **SeniorCitizen:** Indicates if the customer is a senior citizen (1) or not (0).
   b) **Service Details:**
      - **PhoneService:** Whether the customer has phone service (Yes or No).
      - **MultipleLines:** Indicates if the customer has multiple phone lines (Yes, No, or No phone service).
      - **InternetService:** Type of internet service (DSL, Fiber optic, or No).
      - **OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies:** Services subscribed by the customer (Yes, No, or No internet service).
   c) **Account Information:**
      - **Contract:** Type of customer contract (Month-to-month, One year, or Two year).
      - **PaperlessBilling:** Whether billing is paperless (Yes or No).
      - **PaymentMethod:** Customer's payment method (Electronic check, Mailed check, Bank transfer, or Credit card).
      - **MonthlyCharges:** Monthly charges paid by the customer (numerical).
      - **TotalCharges:** Total amount charged to the customer (numerical).
      - **Tenure:** Number of months the customer has stayed with the company (numerical).
   d) **Churn Status (Target Variable):**
      - **Churn:** Whether the customer has churned (Yes or No).

IV. EXPERIMENTS AND RESULTS

1) **Feature Analysis** Feature analysis in our project focused on understanding the relationships between the dataset's features and their impact on the target variable (Churn). Below is a summary of the key steps and insights gained during the feature analysis phase:

- **Correlation Analysis**
  - **Objective:** Identify relationships between numerical features and their influence on churn.
  - **Approach:**
    - A correlation matrix was generated for numerical variables such as Tenure, MonthlyCharges, and TotalCharges.
    - Heatmaps were used to visualize these relationships.

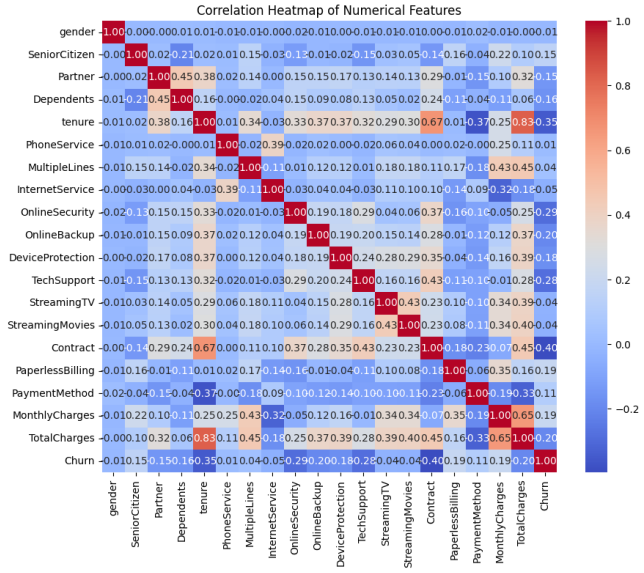

Fig. 1. Correlation Heatmap

    - **Insights:**
      a) Tenure showed a negative correlation with churn, indicating customers with longer tenures were less likely to churn.
      b) MonthlyCharges had a slight positive correlation with churn, suggesting higher charges might contribute to customer attrition.
      c) TotalCharges exhibited a complex relationship due to its dependency on both Tenure and MonthlyCharges.

- **Multicollinearity Check**
  - **Objective:** Ensure features are independent or minimally correlated with each other to avoid redundancy.
  - **Approach:**
    a) Variance Inflation Factor (VIF) was calculated to detect multicollinearity.
    b) Features with high VIF values were flagged for potential removal.

```
-----------------------------------------
Variance Inflation Factor Analysis:
-----------------------------------------

          Feature       VIF
0   SeniorCitizen  1.256364
1          tenure  2.617403
2  MonthlyCharges  2.924996
```

Fig. 2. Variance Inflation Factor Analysis

  - **Insights:**
    a) **MonthlyCharges:**
      - **VIF:** 2.92 suggests mild multicollinearity.
      - In the correlation heatmap, MonthlyCharges shows a **strong correlation with TotalCharges (0.65)**, which could indicate some dependency. However, since TotalCharges is not included in the VIF analysis (likely removed due to its high correlation with tenure and MonthlyCharges), the multicollinearity here seems manageable.
      - **tenure:**
        - **VIF:** 2.62 indicates slight multicollinearity but still within acceptable levels (¡5).
        - The heatmap shows tenure correlating moderately with TotalCharges (0.45) and negatively with Churn. This is expected since TotalCharges depends on both MonthlyCharges and tenure.
      - **SeniorCitizen:**
        - **VIF:** 1.25 indicates minimal multicollinearity.
        - In the heatmap, SeniorCitizen shows weak correlations with all features, confirming it is relatively independent.

2) **Visualization Strategies**

a) **Dimensionality Reduction:** The visualization strategies used in this project provided valuable insights into the relationships between features, data structure, and customer churn behavior.

- **PCA:** Captured 85% of the variance with the first two components, which shows significant overlap between "Churn" and "No Churn" groups, indicating limited separability in the first two principal components, but it helps us identify variance captured by these features.
- **t-SNE and UMAP:** t-SNE and UMAP revealed slight separations between churn and non-churn clusters, indicating underlying patterns in the data. It helped visualize customer behavior in a reduced feature space and validate the impact of key features on churn.

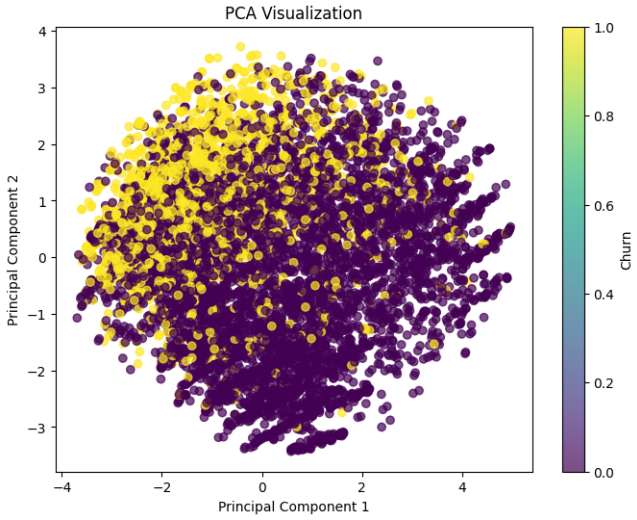

Fig. 3. PCA Visualization

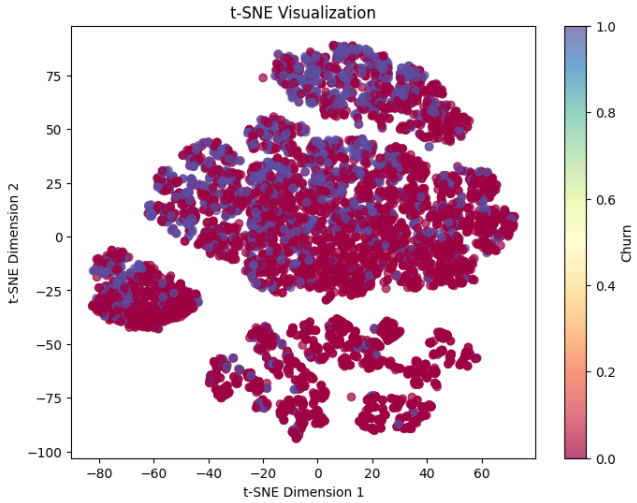

Fig. 4. t-SNE Visualization

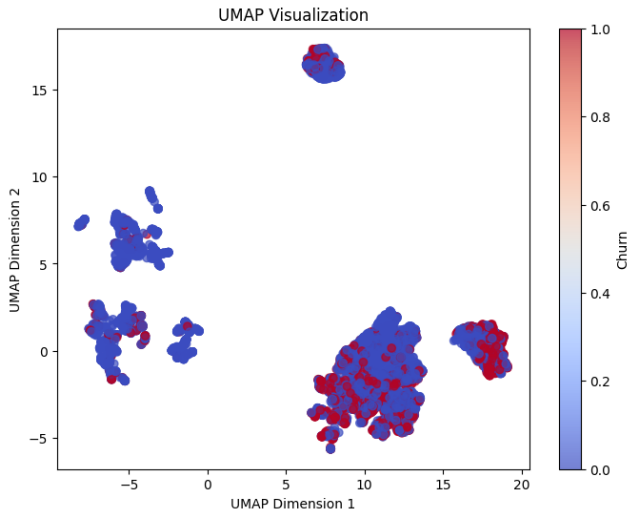

Fig. 5. UMAP Visualization

3) **Unsupervised Learning Techniques**

a) **K-Means Clustering**
Process:

- **Preprocessing:**
  - Categorical features were encoded using LabelEncoder.
  - Features were scaled using StandardScaler to ensure uniformity across dimensions.
- **Implementation:**
  - Applied K-Means with n-clusters=3 and random-state=42.
  - Visualized clusters using a scatter plot of the first two principal components or features.
- **Insights:**
  - Effective for identifying clusters in structured data.
  - Challenges included sensitivity to outliers and the need to predefine the number of clusters.

b) **DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Process:**

- **Preprocessing:**
  - Like K-Means, features were scaled using StandardScaler.
- **Implementation:**
  - DBSCAN with eps=1.5 (maximum neighborhood radius) and min-samples=10 (minimum points in a cluster).
  - Visualized clusters and identified noise points (labeled as -1).
- **Insights:**
  - Successfully detected clusters of arbitrary shapes and handled outliers effectively.
  - Highly sensitive to hyperparameters, particularly eps and min-samples.

4) **Cross-Validation Results Strategy Used:** Stratified K-Fold Cross-Validation and K-Fold Cross-Validation were applied to ensure a balanced representation of the target class (Churn) across folds due to potential imbalance in the dataset.

a) **Stratified K-Fold Cross-Validation Results**

- **Fold Accuracies:**
  - Range: 0.7763 - 0.8098
  - Mean Accuracy: 0.79
- **Fold F1 Scores:**
  - Range: 0.5116 - 0.5988
  - Mean F1 Score: 0.56
- **Insights:**
  i) **Balanced Representation:** Stratified K-Fold ensures the target variable (Churn) is proportionately distributed across folds, making it suitable for imbalanced datasets.

ii) **Higher Accuracy Variance:** The accuracy variance between folds indicates slight instability in performance. This might be due to data noise or uneven feature distributions within the folds.

iii) **F1-Score Limitation:** The F1-score reflects moderate performance. The model struggles to balance precision and recall, possibly due to the imbalanced target class.

```
----------------------------------------
Stratified K-Fold Cross-Validation
----------------------------------------

Fold Accuracy: 0.8098, F1 Score: 0.5988
Fold Accuracy: 0.7970, F1 Score: 0.5806
Fold Accuracy: 0.7977, F1 Score: 0.5714
Fold Accuracy: 0.7763, F1 Score: 0.5116
Fold Accuracy: 0.7770, F1 Score: 0.5184
```

Fig. 6. Result

b) **K-Fold Cross-Validation Results**

- **Fold Accuracies:**
  - Range: 0.7827 - 0.8084
  - Mean Accuracy: 0.79
- **Fold F1 Scores:**
  - Range: 0.5174 - 0.6064
  - Mean F1 Score: 0.56
- **Insights:**
  i) **No Stratification:** K-Fold does not stratify the data, which could lead to imbalanced target class distributions in folds. Despite this, the accuracies and F1 scores are comparable to those of Stratified K-Fold, indicating minimal impact in this case.
  ii) **Higher Variance in F1 Scores:** The F1-score range is slightly broader than in Stratified K-Fold. This could imply greater inconsistency in identifying the minority class (Churn).

```
----------------------------------------
K-Fold Cross-Validation
----------------------------------------

Fold Accuracy: 0.7970, F1 Score: 0.5545
Fold Accuracy: 0.8084, F1 Score: 0.6064
Fold Accuracy: 0.7913, F1 Score: 0.5664
Fold Accuracy: 0.7940, F1 Score: 0.5497
Fold Accuracy: 0.7827, F1 Score: 0.5174
```

Fig. 7. Result

5) **Simple Model Training and Validation**

a) **Hyperparameter Tuning**

- **Model Trained:** Random Forest Classifier.
- **Best Hyperparameters Identified:**
  - max-depth: 10
  - min-samples-split: 5
  - o n-estimators: 50
- **Tuning Process:**
  - Hyperparameters were optimized using Grid Search (or another technique) and evaluated on a validation set.
  - Validation accuracy and F1-score were the guiding metrics.

b) **Validation Performance**

- **Metrics on Validation Set:**
  - Accuracy: **0.7942**
  - F1-Score: **0.5685**
  - ROC-AUC: **0.7637**
- **Insights:**
  - The model performs well on the majority class (No Churn) with a high precision and recall.
  - Performance for the minority class (Churn) is significantly lower, with an F1-score of 0.57, highlighting challenges with the imbalanced dataset.
  - The weighted F1-score across both classes is 0.79, reflecting the imbalance.

c) **Cross-Validation Performance**

- **Cross-Validation F1 Scores:**
  - F1-Scores across folds: **[0.5842, 0.5845, 0.5942, 0.5516, 0.5688]**
  - Mean F1-Score: **0.5767**
- **Insights:**
  i) **Consistency Across Folds:** The F1-scores vary slightly across folds, indicating reasonable consistency.
  ii) **Bias Toward Majority Class:** As in the validation set, the model tends to favor the majority class (No Churn) during cross-validation.

```
Best Hyperparameters :  {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 50}

----------------------------------------
Validation Performance :
----------------------------------------

Accuracy : 0.7942
F1 Score : 0.5685
ROC AUC : 0.7637

----------------------------------------
Classification Report :
----------------------------------------

              precision    recall  f1-score   support

           0       0.84      0.90      0.86      1035
           1       0.64      0.51      0.57       374

    accuracy                           0.79      1409
   macro avg       0.74      0.70      0.72      1409
weighted avg       0.78      0.79      0.79      1409


----------------------------------------
Cross-Validation Performance :
----------------------------------------

Cross-Validation F1 Scores : [0.5842026825633383, 0.5845697329376854, 0.5941807044410413, 0.5516173873651772, 0.5688350983358548]
Mean CV F1 Score : 0.5767
```

Fig. 8. Result

6) **Performance Improvement Analysis**

a) **Regularization:** Logistic Regression with L1 Penalty

- **Results:**
  - Accuracy: **0.7963**
  - F1 Score: **0.5786**
  - ROC AUC: **0.7102**
- **Key Insights:**
  - **Reduced Overfitting:** Applying the L1 penalty improved the model's ability to generalize by shrinking less important feature coefficients to zero.
  - **Performance Stability:** The model maintained a slightly improved F1 score compared to the baseline model, indicating better handling of the minority class (Churn).
  - **Trade-offs:** While accuracy and F1 scores improved slightly, the model's ROC AUC score (0.7102) is moderate, indicating room for further improvement.

```
----------------------------------------
Logistic Regression with L1 Regularization :
----------------------------------------

Accuracy : 0.7963
F1 Score : 0.5786
ROC AUC : 0.7102
```

Fig. 9. Result

b) **Gradient Boosting with Early Stopping**

- **Results:**
  - Accuracy: **0.7899**
  - F1 Score: **0.5634**
  - ROC AUC: **0.7008**
- **Key Insights:**
  - **Early Stopping Benefits:** Early stopping helped prevent overfitting during training, leading to a more robust model.
  - **Minor Improvements:** Accuracy and ROC AUC are comparable to the baseline, but the F1 score is slightly lower compared to Logistic Regression with L1 regularization. This suggests the Gradient Boosting model struggles with the minority class.
  - **Model Robustness:** The model performed consistently with no significant overfitting, highlighting the effectiveness of early stopping in controlling complexity.

```
----------------------------------------
Gradient Boosting with Early Stopping:
----------------------------------------

Accuracy: 0.7899
F1 Score: 0.5634
ROC AUC: 0.7008
```

Fig. 10. Result

7) **2 more experiments**

a) **Experiment 1:** Feature Selection Using Recursive Feature Elimination (RFE)

- **Results:**
  - **Selected Features:** 12 out of the original features.
    - Feature indices: [0, 4, 6, 7, 8, 9, 11, 14, 15, 16, 17, 18]
  - **Performance Metrics:**
    - Accuracy: **0.7771**
    - F1 Score: **0.5242**
    - ROC AUC: **0.6767**
  - **Classification Report:**

Class	Precision	Recall	F1-Score	Support
0 (No Churn)	**0.82**	**0.89**	**0.85**	1035
1 (Churn)	**0.60**	**0.46**	**0.52**	374

Fig. 11. Table

- **Insights:**

i) **Improved Interpretability:** Selecting only 12 features reduces model complexity and enhances interpretability, which can be beneficial for understanding key predictors of churn.

ii) **Performance Trade-offs:**
  - While accuracy is slightly lower compared to the baseline, the F1 score for the minority class (Churn) remains moderate.
  - Precision for Churn is relatively good, but recall is lower, indicating the model misses some true positive cases.

iii) **Limitations:** The drop in overall performance metrics, particularly ROC AUC, indicates that feature selection might have removed some relevant features, slightly impacting the model's predictive power.

```
----------------------------------------
Experiment 1: Feature Selection
----------------------------------------

Selected Features Index : [ 0  4  6  7  8  9 11 14 15 16 17 18]
Optimal Number of Features : 12

Performance with Feature Selection :
Accuracy : 0.7771
F1 Score : 0.5242
ROC AUC : 0.6767

Classification Report :
              precision    recall  f1-score   support

           0       0.82      0.89      0.85      1035
           1       0.60      0.46      0.52       374

    accuracy                           0.78      1409
   macro avg       0.71      0.68      0.69      1409
weighted avg       0.76      0.78      0.77      1409
```

Fig. 12. Result of Experiment 1

b) **Experiment 2:** Increasing Model Complexity Using Gradient Boosting

- **Results:**
  - **Performance Metrics:**
    - Accuracy: **0.7899**
    - F1 Score: **0.5634**
    - ROC AUC: **0.7008**
  - **Classification Report:**

Class	Precision	Recall	F1-Score	Support
0 (No Churn)	**0.83**	**0.89**	**0.86**	1035
1 (Churn)	**0.63**	**0.51**	**0.56**	374

Fig. 13. Table

- **Insights:**
  i) **Enhanced Model Power:** Gradient Boosting leverages boosting iterations to improve performance, particularly for the minority class (Churn), achieving a slightly higher F1 score compared to Experiment 1.
  ii) **Balanced Trade-offs:** The model achieves a good balance between precision and recall for both classes, reflected in an improved F1 score and ROC AUC compared to Experiment 1.
  iii) **Computational Complexity:** Gradient Boosting increases training time and resource usage due to the higher number of boosting iterations and deeper trees.
  iv) **Limitations:** While accuracy and F1 scores are better, the recall for Churn remains moderate, indicating that further optimization may be needed to improve sensitivity.

```
----------------------------------------
Experiment 2: Increasing Model Complexity
----------------------------------------


Performance with Gradient Boosting :
Accuracy : 0.7899
F1 Score : 0.5634
ROC AUC : 0.7008

Classification Report :
              precision    recall  f1-score   support

           0       0.83      0.89      0.86      1035
           1       0.63      0.51      0.56       374

    accuracy                           0.79      1409
   macro avg       0.73      0.70      0.71      1409
weighted avg       0.78      0.79      0.78      1409
```

Fig. 14. Result of Experiment 2

CONCLUSION

This project analyzed customer churn in the telecommunications industry using a structured approach encompassing data exploration, visualization, and machine learning modeling. Key insights revealed strong correlations between churn likelihood and features like tenure and contract type, with customers on month-to-month contracts or with lower tenure showing higher churn rates, and payment methods such as electronic checks being associated with elevated churn risk. Visualization techniques, including PCA and UMAP, uncovered separable patterns in customer segments, identifying potential clusters of at-risk customers, while correlation plots helped isolate significant predictors for effective feature selection. In modeling, baseline approaches like logistic regression and decision trees provided initial performance metrics, with hyperparameter tuning further refining results. Cross-validation exposed overfitting in complex models, underscoring the importance of regularization and balanced validation strategies. Challenges such as class imbalance, with fewer churn cases compared to non-churn, were addressed using techniques like SMOTE, and careful attention to feature selection and scaling proved crucial for optimizing model performance.

FUTURE SCOPE

1) **Prediction:** Deploy the model in a real-time environment to proactively identify at-risk customers.
2) **Customer Segmentation:** Further, refine customer clusters to design personalized retention strategies.
3) **Incorporate More Data:** Use additional data sources such as customer support interactions and feedback for deeper insights.
4) **Optimize Models:** Explore ensemble methods and hyperparameter tuning for further performance improvement.

REFERENCES

[1] K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[2] D. Deng, "DBSCAN Clustering Algorithm Based on Density," 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), Hefei, China, 2020, pp. 949-953, doi: 10.1109/IFEEA51475.2020.00199.

[3] T. N. Varunram, M. B. Shivaprasad, K. H. Aishwarya, A. Balraj, S. V. Savish and S. Ullas, "Analysis of Different Dimensionality Reduction Techniques and Machine Learning Algorithms for an Intrusion Detection System," 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), Arad, Romania, 2021, pp. 237-242, doi: 10.1109/ICCCA52192.2021.9666265.

[4] K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 83-87, doi: 10.1109/ICCMC48092.2020.ICCMC-00016.

[5] Y. Xiao, W. Huang and J. Wang, "A Random Forest Classification Algorithm Based on Dichotomy Rule Fusion," 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 2020, pp. 182-185, doi: 10.1109/ICEIEC49280.2020.9152236.

GITHUB LINK

Link: https://github.com/DeePawar28/CSP-571-Group-25