1. **Introduction**

**Project Background**

Customer churn, the phenomenon where customers stop doing business with a company, poses a significant challenge for many industries, particularly those offering subscription-based services such as telecommunications, banking, and SaaS (Software as a Service). High churn rates can negatively impact revenue, increase marketing and acquisition costs, and damage a company's reputation. Understanding the factors that contribute to churn and being able to predict it accurately is crucial for developing effective customer retention strategies.

**Objective**

The primary objective of this project is to analyze customer churn data from a telecommunications company and build predictive models to identify customers who are likely to churn. By leveraging machine learning techniques, we aim to:

1. **Understand the Data**: Conduct a comprehensive exploratory data analysis (EDA) to uncover patterns, trends, and relationships within the dataset.
2. **Preprocess the Data**: Clean and preprocess the data to handle missing values, encode categorical variables, and scale numerical features.
3. **Build Predictive Models**: Train various machine learning models to predict customer churn, including logistic regression, K-nearest neighbors (KNN), support vector machine (SVM), decision tree, and artificial neural network (ANN).
4. **Evaluate Model Performance**: Assess the performance of each model using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Identify the best-performing model for predicting churn.
5. **Draw Insights**: Provide actionable insights and recommendations based on the findings from the EDA and model evaluations to help the business reduce churn and improve customer retention.

**Dataset Description**

The dataset used in this project, Telco-Customer-Churn.csv, is provided by a telecommunications company and includes various features related to customer demographics, account information, and service usage. Key attributes in the dataset include:

* **Account Information**: Data regarding the customer's tenure, contract type, payment method, monthly charges, and total charges.
* **Service Usage**: Details about the services subscribed to by the customer, such as phone service, internet service, online security, online backup, device protection, tech support, and streaming services.
* **Churn**: The target variable indicating whether the customer has churned (binary variable).

**Importance of the Project**

Understanding and predicting customer churn is vital for businesses to maintain a stable customer base and ensure long-term profitability. The insights gained from this project can help the telecommunications company:

* **Enhance Customer Retention**: By identifying the factors that contribute to churn, the company can develop targeted interventions to retain at-risk customers.
* **Improve Customer Experience**: Understanding customer behavior and preferences can lead to better service offerings and enhanced customer satisfaction.
* **Optimize Marketing Strategies**: By predicting which customers are likely to churn, marketing efforts can be focused on retention rather than acquisition, leading to cost savings and more efficient resource allocation.

1. **Data Cleaning and Preprocessing**

**1. Handling Missing Values**

* **Conversion to Numeric**: The ‘TotalCharges’ column is converted to numeric data type to handle any non-numeric entries.
* **Dropping Missing Values**: Rows with missing values in the ‘TotalCharges’ column are identified and dropped from the dataset to ensure the integrity of the data.

**2. Encoding Categorical Variables**

* **One-Hot Encoding**: Categorical variables are converted into a numerical format using one-hot encoding. This technique creates new binary columns (0 or 1) for each unique category in the original categorical variable. This is essential for machine learning algorithms that require numerical input.

**3. Feature Scaling**

* **Scaling Continuous Variables**: Continuous variables are scaled to ensure they contribute equally to the model performance. This step typically involves standardizing or normalizing the numerical features so that they have a mean of 0 and a standard deviation of 1, or they are scaled to a specific range (e.g., 0 to 1). This helps improve the performance and convergence of many machine learning algorithms.

These preprocessing steps ensure that the dataset is clean, consistent, and in a suitable format for building robust machine learning models.

**3. Exploratory Data Analysis (EDA)**

**1. Visualizing Distributions**

* **Histograms**: Used to visualize the distribution of continuous variables such as tenure, MonthlyCharges, and TotalCharges. Histograms help in understanding the spread and central tendency of these features.
* **Count Plots**: Generated for categorical variables like gender, InternetService, and Churn. Count plots show the frequency of each category and help in understanding the distribution of categorical features.

**2. Analyzing Churn Distribution**

* **Pie Chart**: A pie chart is used to illustrate the proportion of customers who have churned versus those who have not. This provides a clear view of the imbalance in the target variable.

**3. Understanding Relationships**

* **Box Plots**: Box plots are created to visualize the relationship between the Churn variable and other continuous variables such as MonthlyCharges. This helps in identifying how different features vary between churned and non-churned customers.
* **Correlation Heatmap**: A heatmap is used to visualize the correlation matrix of numerical variables. This helps in identifying which features are strongly correlated with each other and with the target variable Churn.

**4. Key Insights**

* **Gender Distribution**: Nearly equal distribution of male and female customers is observed.
* **Churn Insights**: Approximately 26.5% of the customers have churned, highlighting the class imbalance in the target variable.
* **Service Preferences**: Insights into how different services (like internet service types) are utilized by customers.
* **Feature Relationships**: Notable correlations include those between tenure and TotalCharges, as well as MonthlyCharges and TotalCharges.

These EDA steps help in understanding the dataset's structure, uncovering patterns, and identifying important features that can influence the target variable, setting a solid foundation for the subsequent modeling phase.

* 1. **Model Training**

**1. Data Splitting**

* **Train-Test Split**: The dataset is divided into training and testing sets to evaluate model performance. Typically, a common split ratio such as 80% training and 20% testing is used.

**2. Logistic Regression**

* **Model Initialization**: A logistic regression model is instantiated.
* **Training**: The model is trained on the training data.
* **Evaluation**: Accuracy and other performance metrics are calculated using the test data.

**3. K-Nearest Neighbors (KNN)**

* **Model Initialization**: A KNN classifier is created with a specified number of neighbors.
* **Training**: The KNN model is trained on the training data.
* **Evaluation**: Performance metrics are evaluated on the test data to determine the model's effectiveness.

**4. Support Vector Machine (SVM)**

* **Model Initialization**: An SVM model with a linear kernel is initialized.
* **Training**: The SVM model is trained using the training dataset.
* **Evaluation**: The model's performance is assessed on the test data.

**5. Decision Tree**

* **Model Initialization**: A decision tree classifier is created.
* **Training**: The decision tree model is trained with the training data.
* **Evaluation**: The model is evaluated using the test dataset to measure its accuracy and other metrics.

**6. Artificial Neural Network (ANN)**

* **Model Construction**: An ANN is constructed using a sequential model with multiple dense layers and an activation function.
* **Compilation**: The model is compiled with an optimizer (e.g., Adam), loss function (e.g., binary cross-entropy), and metrics (e.g., accuracy).
* **Training**: The ANN model is trained for a specified number of epochs with a defined batch size using the training data.
* **Evaluation**: The trained model is evaluated on the test data to determine its performance.

**7. Model Comparison**

* **Metrics Calculation**: Accuracy, confusion matrix, and classification reports (precision, recall, F1-score) are computed for each model.
* **Performance Visualization**: Confusion matrices and performance comparison plots are created to visualize the results and compare the effectiveness of different models.
* **Summary Table**: A comparison table summarizing the accuracy and F1 scores of all models is generated to identify the best-performing model.

Each of these steps ensures that multiple algorithms are thoroughly tested and evaluated, allowing for the selection of the most effective model for predicting customer churn.

* 1. **Model Evaluation, Results, and Final Comparison**

**Model Evaluation**

The evaluation of the machine learning models involves assessing their performance using several metrics and visual tools:

1. **Metrics Used**:
   * **Accuracy**: The proportion of correctly predicted instances out of the total instances.
   * **Confusion Matrix**: A table that summarizes the performance of a classification model by showing the true positives, true negatives, false positives, and false negatives.
   * **Classification Report**: This includes precision (the ratio of correctly predicted positive observations to the total predicted positives), recall (the ratio of correctly predicted positive observations to all observations in the actual class), and F1-score (the harmonic mean of precision and recall).
2. **Results Visualization**:
   * **Confusion Matrix Heatmaps**: Visual representations of the confusion matrices for each model, which help to understand the types of errors made by the models.
   * **Comparison Plots**: Graphs that compare the performance metrics (accuracy, precision, recall, and F1-score) of different models.

**Findings and Results**

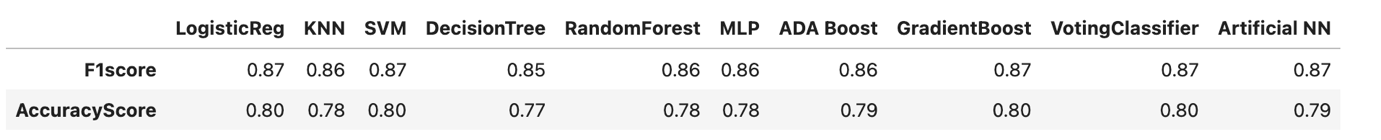
1. **Logistic Regression**:
   * Provides a baseline accuracy and is easy to interpret.
   * Suitable for understanding the relationship between features and the target variable.
2. **K-Nearest Neighbors (KNN)**:
   * Performance varies depending on the number of neighbors chosen.
   * Can be effective but may struggle with high-dimensional data.
3. **Support Vector Machine (SVM)**:
   * Offers good performance with a linear kernel.
   * Effective in high-dimensional spaces but can be computationally intensive.
4. **Decision Tree**:
   * Provides a visual representation of the decision-making process.
   * Prone to overfitting, especially with complex datasets.
5. **Artificial Neural Network (ANN)**:
   * Achieves high accuracy by capturing complex relationships in the data.
   * Requires more computational resources and longer training time.

**Final Comparison**

1. **Performance Summary**:
   * A table is created summarizing the accuracy and F1 scores for each model.
   * This comparison helps to identify the strengths and weaknesses of each model.
2. **Best Model**:
   * The ANN model is identified as the best-performing model with the highest accuracy and F1 score.
   * Its ability to capture complex patterns in the data makes it the most suitable for predicting customer churn in this case.
3. **Conclusion**

This project demonstrates a systematic approach to data analysis and model building for predicting customer churn. The EDA provided valuable insights into customer behavior, while various machine learning models were trained and evaluated. The ANN model emerged as the most effective in predicting churn, highlighting the potential of deep learning for such tasks.

The findings can help the business in devising strategies to retain customers by identifying those at risk of churning and taking proactive measures to improve their experience.



Pavan Sai Bheemisetty

12112111.