Customer Churn Prediction in Telecommunications Company

This report explores customer churn in a telecommunications dataset. Customer churn refers to customers leaving a service provider for a competitor. The goal is to understand the factors influencing churn and develop models to predict which customers are at risk.

Data Exploration

The project starts by gathering the necessary tools for data analysis (pandas, numpy), visualization (matplotlib, seaborn), data cleaning (scikit-learn's LabelEncoder), and model building. The data is loaded from a CSV file and stored in a structured format for analysis.

We then get a general idea of the data by examining the data types and the presence of missing values in each column. Additionally, we explore the distribution of the churn labels (Yes/No) to understand the balance between customers who churn and those who don't.

Some data columns contain categorical information (e.g., text labels). To prepare this data for machine learning models, these categories are converted into numerical values using a technique called label encoding.

Finally, we perform a correlation analysis to see how features in the data relate to customer churn. This can help identify which factors might be most important for predicting churn.

Model Building and Evaluation

1. **Feature Selection and Preprocessing:**

When building models, it's important to focus on features most relevant to the task. For instance, we might exclude features like customer ID or gender from the model as they likely don't directly relate to churn. The target variable (churn label) is separated for model training.

Numerical features in the data are often scaled to a standard range for better model performance. This helps ensure all features contribute equally during model training.

2. **Train-Test Split:**

The data is divided into training and testing sets. The training set (typically the larger portion) is used to train the models, while the testing set is used to evaluate their performance on unseen data. This helps prevent overfitting, where a model performs well on the training data but poorly on new data.

3. **Machine Learning Models:**

Several machine learning models are explored to predict customer churn, including:

```
* **K-Nearest Neighbors (KNN):** This model classifies new data points
based on the labels of their nearest neighbors in the training data.  The
number of neighbors considered (k) can be tuned to optimize performance.
```

* **Logistic Regression:** This model estimates the probability of an event (like customer churn) occurring based on its features.

* **Decision Tree Classifier:** This model creates a tree-like structure to classify data points based on a series of decision rules.

* **Support Vector Machine (SVM):** This model creates a hyperplane that separates data points belonging to different classes (churn/no churn) with the largest possible margin.

Each model is trained on the training data and then evaluated on the testing set. Metrics like accuracy and classification reports are used to assess how well the models perform in identifying customers who are likely to churn.

Conclusion

This project explored a customer churn dataset and built various models to predict churn. The KNN model achieved the highest accuracy on the test set. However, there's always room for improvement. Future work could involve incorporating additional features, exploring different models, and fine-tuning parameters to potentially achieve even better results.

The insights gained from this project can be used by the telecommunications company to develop targeted strategies to retain customers and reduce churn rates. This can lead to increased customer satisfaction, loyalty, and ultimately, higher profitability.