LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

Project : *CUSTOMER CHURN PREDICTION*

BY GUMMALA CHANDRAKANTH

Project Report:

1. Introduction

The objective of this project was to predict customer churn in a telecommunications company using a variety of features related to customer demographics, services subscribed, and account information. Churn prediction helps businesses to identify customers likely to leave and take proactive measures to retain them.

2. Exploratory Data Analysis (EDA)

Data Overview

The dataset contains 100 columns including customer information such as gender, senior citizen status, partner and dependents status, tenure, services subscribed (phone, multiple lines, internet, online security, online backup, device protection, tech support, streaming TV, and streaming movies), contract type, paperless billing, payment method, monthly charges, total charges, and churn status.

Key Findings

- **Churn Distribution**: Approximately 27% of customers in the dataset have churned, indicating an imbalance that needs to be addressed during model evaluation.

- **Service Subscriptions**: A significant portion of churned customers are those with fiber optic internet service. Customers without internet service have a lower churn rate.
- **Contract Type**: Customers with month-to-month contracts are more likely to churn compared to those with one-year or two-year contracts.
- **Monthly Charges**: Higher monthly charges are associated with a higher churn rate. This indicates that cost may be a factor in the decision to churn.
- **Tenure**: Customers with shorter tenures are more likely to churn. Long-term customers show higher loyalty.

3. Feature Engineering

Handling Missing Values

- Missing values in the `TotalCharges` column were filled with the mean of the column.

Encoding Categorical Variables

- Categorical variables were encoded using `LabelEncoder` for binary

Final Feature Set

The final feature set included variables such as gender, senior citizen status, partner status, dependents status, tenure, various service subscriptions, contract type, paperless billing, payment method, monthly charges, and total charges.

4.MODELS USED:

1.Logistic Regression

Evaluation Results

- **Accuracy**: 0.75
- **Precision**: 0.82
- **Recall**: 0.82
- **F1 Score**: 0.81

2.Random Forest

Evaluation Results

- **Accuracy**: 0.81
- **Precision**: 0.82
- **Recall**: 0.82
- **F1 Score**: 0.81

3 .XGBoost

Evaluation Results

- **Accuracy**: 0.86
- **Precision**: 0.86
- **Recall**: 0.86
- **F1 Score**: 0.86

4. Decision Tree

Evaluation Results

- **Accuracy**: 0.87
- **Precision**: 0.87
- **Recall**: 0.87
- **F1 Score**: 0.87

The Decision Tree classifier provided reasonable results among all

5. Challenges Faced

- **Class Imbalance**: The dataset had an imbalanced distribution of churn vs. non-churn cases. Techniques such as resampling or using evaluation metrics like F1-score, which balances precision and recall, were necessary.

- **Feature Selection**: Deciding which features to include in the model was crucial. Including too many features can lead to overfitting, while too few can miss important information.

- **Handling Categorical Variables**: The dataset contained numerous categorical variables that required careful encoding to ensure the model could interpret them correctly.

6. Conclusion

The Decision Tree classifier provided reasonable results, but there is room for improvement, particularly in improving recall. Future steps could include experimenting with more sophisticated models such as Random Forests or Gradient Boosting Machines, and techniques to handle class imbalance more effectively, such as SMOTE (Synthetic Minority Over-sampling Technique) or class weighting.

Through this project, we gained valuable insights into the factors affecting customer churn and developed a model that can help in identifying at-risk customers, enabling the company to take proactive measures for customer retention.