**Logistic Regression**

**Vs.**

**Decision Tree/Random Forest**

#### **Problem Statement**

The goal of this project is to predict whether a customer will switch to another telecom provider (churn) using a dataset with 21 variables related to customer behavior, such as monthly bill, internet usage, etc.

#### **Steps Performed**

##### **Logistic Regression**

1. **Data Cleaning and Preprocessing**
   * Handled missing values.
   * Encoded categorical variables.
   * Normalized numerical features.
2. **Train-Test Split**
   * Split the data into training and testing sets.
3. **Feature Scaling**
   * Standardized features to ensure they are on a similar scale.
4. **Feature Selection**
   * Removed highly correlated features.
   * Used Recursive Feature Elimination (RFE) to select the most significant features.
5. **Handling Multicollinearity**
   * Calculated Variance Inflation Factors (VIFs) to detect multicollinearity.
   * Removed features with high VIFs and rebuilt the model.
6. **Model Building**
   * Built and trained the logistic regression model.
7. **Model Evaluation**
   * Calculated accuracy, recall, and precision.
   * Plotted the ROC curve and calculated the Area Under the Curve (AUC).
   * Logistic Regression Accuracy: 78.3%
   * Logistic Regression AUC (Area under the ROC curve): 0.85
8. **Prediction**
   * Made predictions on the test data.

##### **Decision Trees**

1. **Train-Test Split**
   * Split the data into training and testing sets.
2. **Model Building**
   * Built and trained the decision tree classifier.
3. **Hyperparameter Tuning**
   * Used Grid Search Cross-Validation to find the best parameters.
4. **Model Evaluation**
   * Calculated accuracy.
   * Plotted the ROC curve and calculated the AUC.
   * Decision Tree Accuracy: 79.8%
   * Decision Tree AUC (Area under the ROC curve): 0.87

##### **Random Forest**

1. **Train-Test Split**
   * Split the data into training and testing sets.
2. **Model Building**
   * Built and trained the random forest classifier.
3. **Hyperparameter Tuning**
   * Used Grid Search Cross-Validation to find the best parameters.
4. **Model Evaluation**
   * Calculated accuracy.
   * Plotted the ROC curve and calculated the AUC.
   * Random Forest Accuracy: 80.5%
   * Random Forest AUC (Area under the ROC curve): 0.94

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic Regression** | **Decision Tree** | **Random Forest** |
| **Accuracy** | 78.3% | 79.8% | 80.5% |
| **AUC** | 0.85 | 0.87 | 0.94 |

#### **Conclusion**

Random Forests provided a significant improvement in results compared to both logistic regression and decision trees, with much less effort. They leveraged the predictive power of decision trees and learned much more than a single decision tree could do alone. However, Random Forests lack visibility regarding key features and the direction of their effects, which logistic regression handles well. If interpretation is not of key significance, Random Forests are highly effective.