Introduction to Data Mining
Final Report

Customer Churn Analysis in Telecom Industry

Contents

Choosing Machine Learning Algorithms & Explaining

For customer churn analysis, we considered several machine learning algorithms based on our datasets characteristics and the problem we had. These are the algorithms that used mostly for binary classification problems like our customer churn prediction problem:

1. Logistic Regression
2. Decision Trees
3. Random Forest

Advantages and Disadvantages of Selected Algorithms

Logistic Regression

- **Advantages:** Logistic Regression is known for its interpretability and efficiency. The model provides clear and interpretable results, making it well-suited for scenarios where understanding the impact of each feature is crucial. Moreover, it is computationally efficient, making it suitable for large datasets and real-time applications.
- **Disadvantages:** One limitation of Logistic Regression is its assumption of linearity between independent variables and the log-odds of the dependent variable. While this assumption may not always hold, it is often reasonable for many real-world problems.

Decision Tree

- **Advantages:** Decision Trees excel in modeling non-linear relationships and do not assume linearity between features. They offer interpretability through visual representations, making them suitable for scenarios where understanding the decision-making process is important.
- **Disadvantages:** However, Decision Trees are prone to overfitting, especially when deep. This can be mitigated through hyperparameter tuning and pruning. Additionally, they may exhibit instability, producing different tree structures with small changes in the data.

Random Forest

- **Advantages:** Random Forest, an ensemble of Decision Trees, addresses the overfitting concern associated with individual trees. It is robust to outliers and noise in the data, providing improved generalization performance. The model also offers feature importance ranking, aiding in identifying influential features.
- **Disadvantages:** On the downside, Random Forest models can be computationally intensive. While they provide feature importance, interpreting the combined effect of features in the ensemble can be challenging. Nevertheless, the trade-off is often justified for enhanced predictive accuracy.

In summary, Random Forest is a compelling choice when robust predictive performance and versatility are paramount. Therefore, we have opted for this algorithm as our foundational model and proceeded with subsequent steps based on its capabilities.

Definitions and Parameters

Logistic Regression

Logistic Regression is a statistical model that analyzes a dataset to predict the probability of a binary outcome. It models the relationship between one dependent binary variable and one or more independent variables, providing probabilities in the range [0, 1]. These are the parameters to tune:

- **C:** Controls the trade-off between fitting the training data and preventing overfitting.
- **Solver:** The algorithm to use in the optimization problem (e.g., 'liblinear', 'newton-cg').
- **Penalty:** Type of regularization ('l1' or 'l2').
- **Max Iterations:** Maximum number of iterations for the solver to converge.

Decision Tree

A Decision Tree is a tree-like model where each node represents a decision based on input features. It recursively splits the data into subsets, making decisions at each node, until a stopping criterion is met. These are the parameters to tune:

- **Max Depth:** Maximum depth of the tree, controlling the maximum number of levels.
- **Min Samples Split:** Minimum number of samples required to split an internal node.
- **Min Samples Leaf:** Minimum number of samples required to be at a leaf node.
- **Criterion:** The function to measure the quality of a split ('gini' or 'entropy').

Random Forest

Random Forest is an ensemble learning method that constructs a multitude of Decision Trees during training. It outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. These are the parameters to tune:

- **Number of Trees (n_estimators):** The number of trees in the forest.
- **Max Features:** The maximum number of features to consider for the best split.
- **Max Depth:** Maximum depth of the individual trees.
- **Min Samples Split:** Minimum number of samples required to split an internal node.

Base Evaluation Technique

Normally if the dataset is balanced, accuracy could be a good starting point for a quick overview. But if the dataset is imbalanced, combination of precision, recall, and F1-score might be the better choice. Since the outcome we want to achieve is to “churn” or “not churn” of the customer (binary problem), accuracy will serve us well. However, to be sure, we will also include the confusion matrix in our evaluation criteria.

Best Evaluation Techniques

We used GridSearch cross validation technique on both Logistic Regression and Decision Tree. We couldn’t use GridSearch on RandomForest because RandomForest is a very complex algorithm that takes too much to compute so we took a randomized approach. We used RandomSeach cross validation instead of GridSearch cross validation. Here are the results:

1. **Random Forest - Accuracy: 0.7969**

   Best Parameters: {'n_estimators': 50, 'min_samp_split': 10, 'max_feat': 'log2', 'max_depth': 10}

2. **Logistic Regression - Accuracy: 0.7884**

   Best Parameters: {'C': 10, 'max_iter': 300, 'penalty': 'l2', 'solver': 'sag'}

3. **Decision Tree - Accuracy: 0.7656**

   Best Parameters: {'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}

The confusion matrix results are on the jupyter file with visualization along with other datas and visualizations. But in the end wee see that the best performing algorithm is RandomForest with LogisticRegression is little bit behind and the DecisionTree is the worse performing one. Altough all the algorithms performed very closely.

Further Performance Improvement

We used LimeTabularExplainer and Decision Tree visualization methods to develop our interpretation of the model. By interpreting individual forecasts, the LimeTabularExplainer helped ensure local interpretability and contributed to a better understanding of model behavior. In a particular example, understanding the importance of the item and gaining insight into the decision-making process was particularly valuable.

Visualizing the decision tree further enhanced our interpretive efforts by representing the basic structure of our model. This visualization allowed for a sense of how different factors contribute to simple and intuitive decision-making. Such a visualization facilitated a student-friendly understanding of the workings of our model.

Using these techniques, we sought to bridge the gap between complex modeling and interpretation, and provide a clearer understanding of the predictive mechanisms at play This approach is the student friendly, practical adapting model research and development.

Inference

As a result of the study, we observed that some features have a greater effect on people churn, and the following comments can be made based on the outputs we received from the model and visualized using the lime tabular explainer:

-Users with short contract periods are likely to unsubscribe.

-Those who subscribe with their relatives are less likely to leave.

-People who have been churn for many years tend to continue their subscriptions.

Our aim in this project was to provide predictions to telecommunication companies about the situations in which customers may unsubscribe. At this point, we have observed that the model we trained can provide information about whether the people presented to it will unsubscribe or not, with an accuracy rate approaching 80%.

The most difficult part of this project was dealing with corrupt and missing data.