**Churn Prediction Using Decision Tree Algorithm**

Project link

<https://colab.research.google.com/drive/1E9yC6VMnell74zEZlPNxBaFoClAKDxmu#scrollTo=_bAjRQdY7PJ4>

https://colab.research.google.com/drive/1E9yC6VMnell74zEZlPNxBaFoClAKDxmu?usp=sharing

Introduction

This document shows algorithm used to predict behaviour to retain customers. Focus is to predict customer behaviour based on previous customer data. The raw data contains 7043 rows (customers) and 21 columns (features). The “Churn” column is target. Data is available in <https://www.kaggle.com/blastchar/telco-customer-churn>.

Algorithm

1. Data Preparation

Before actual implementation exploratory data analysis done on the data set.

As a part of pre-processing implemented categorical string data to one hot encoded

numeric data.

* Dictionary preparation is done for each column
* Dropped data frame which are empty.
* Fed this dictionary to onehotencoder class:, return column data frame with encoded numeric data. Thus converted from categorical string to numeric data frame.

1. Data split

split the data for train and validation set based on user input. Method ‘train\_test\_split’ takes 2 parameter dataframe and split size (default split size is 10%).This method split the entire data frame to train data frame and test data frame. As test data frame extracted from the total data set helps to validate the prepared model accuracy.

1. Entropy

If we have c different value , entropy is calculated using formula where

* n is continuous value to calculate entropy
* Pi is the probability of occurrence of ith value

1. Information Gain and Build

Is how important a given attribute of the feature vectors. Formula used to calculate information gain is entropy(parent) – [average entropy(children)].

Recursively identified the best split of the data set. Based on threshold value segregated parent ,left and right child node calculated information gain for each node if the gain is grater than previous , new node consider for splitting. Best split method gold the information on left node, right node, threshold ,gain and feature index where it is split.

During build face, best split is identified recursively and gathered node information on split.

1. Model Evaluation

Data split is done for entire data to train and test .Test data set considered for evaluating the model accuracy. Accuracy is calculated for the model by calculating mean value for actual and predicted data which are in comparable with the value. Function is defined by name ‘accuracy’ to calculate prepared model accuracy.

1. Experiment Design and Evaluation
2. Evaluation Results
3. Conclusion