upGrad
INSOFE
INSPIRE • EDUCATE • TRANSFORM

DSC 7456 - Machine Learning:

Objective: Understand the application of machine learning algorithms in real world scenarios and implement the classification techniques taught as part of the module DSC 7456 Machine Learning, to solve the near real time use case from telecommunications domain.

Brief Learning Outcomes:

- Be able to understand, explore Data and identify the dependent and independent variables.
- Be able to merge the Data, Since the realtime data may come from multiple sources.
- Be able to prepare and preprocess the data for modelling .
- Be able to implement advanced machine learning models, analyze and interpret the output.
- Be able to get best features out of all the given attributes.
- Be able to choose an evaluation metric for the problem at hand.
- Be able to improve the model performance by hyper parameter tuning techniques.
- Be able to choose the final best models of all the models.
- Be able to well understand real time bussiness perspectives & build ideas for the constructive growth of a bussiness.

Project - Domain: Telecommunications

Background : Most telecom companies suffer from the customer churn. Churn rate has a strong impact on the life time value of the customer because it affects the length of service and the future revenue of the company. As telecom companies spend lot of money to acquire a new customer and when that customer leaves, the company not only loses the future revenue from that customer but also the resources spent to acquire that customer. Therefore, it is very important for telecom provider to find out potential churn customers, to be able to target-serve them and retain them as much as possible. But Most telecom companies suffer from the customer churn. Churn rate has a strong impact on the life time value of the customer because it affects the length of service and the future revenue of the company. As telecom companies spend lot of money to acquire a new customer and when that customer leaves, the company not only loses the future revenue from that customer but also the resources spent to acquire that customer. Therefore, it is very important for telecom provider to find out potential churn customers, to be able to target-serve them and retain them as much as possible.

Problem Statement: A telecom company has the following data related to its subscribers. It wants to find out potential churn customers to find out the potential churners and to turn them into regular customers by retaining them. Data contains attributes that corresponds to certain reason for each attribute such as:

Cust_Id	: An Unique value for each Customer.
International.Plan	: Whether the customer has International plan or not.
Voice.Mail.Plan	: Whether the customer has Voice Mail plan or not.
X..Vmail.Messages	: Number of Voicemail messages customer has sent.
Total.Day.Minutes	: Total call duration taken by customer per Day in minutes.
Total.Day.Calls	: Number of calls in total customer did per Day.

Total.Day.Charge : Total Charges accordingly per Day.
Total.Eve.Minutes : Total Number of minutes taken by customer in the Evening.
Total.Eve.Calls : Total Number of calls made by customer in the Evening.
Total.Eve.Charge : Total Charges accordingly per Evening calls.
Total.Night.Minutes : Total Number of minutes taken by customer in the Night.
Total.Night.Calls : Total Number of calls made by customer in the Night.
Total.Night.Charge : Total Charges accordingly per Night calls.
Total.Intl.Minutes : Total Number of minutes taken by customer for International calls.
Total.Intl.Calls : Total Number of International calls made by customer.
Total.Intl.Charge : Total Charges accordingly per International calls.
X..customer.Service.Calls : Number of calls made by customer.
Churn : Whether that customer churn out of the company or not.
Trainrows : For Splitting Data into Train & Test sets.

Our Aim is to protect our customers without churning from our company. It can be achieved through best model building based on best features using various machine learning algorithms and get good insights from the data which can help improve company's business.

Steps for Achieving Aim:

1. Reading & Data Preprocessing
   a. Read the Data which is splitted into 4 chunks i.e., Telco_Customer_Call_Details1, Telco_Customer_Call_Details2, Telco_Churn_Details1, Telco_Churn_Details2.
   b. Combine those 4 Datasets into one Dataset.
   c. Explore and understand the data.
   d. Spot the outliers & Handle them properly if any occurred.
   e. Check and handle missing values if any.
   f. Perform type-conversion for required columns.
   g. Perform Standardization & Encodings.
2. Draw good insights from the visualizations. While visualizing choose appropriate graphs.
3. Split the data into train and validation sets based on the column 'trainrows'. Rows with 'yes' belong to train data, those with 'no' belong to validation data.
4. Model Building –
   - Model1 – Basic:
   i. Build Basic model on all attributes, analyze model performance.
   ii. Build the following Machine Learning models:
      a. Decision Trees.
      b. Randomforest.
      c. Adaboost.
      d. Gradient Boosting Machines.
      e. Support Vector Machines.
   iii. Interpret the model output : Significance of overall model.
   iv. Choose the appropriate evaluation metric.

5. Steps to improve Model performance –

(a) Build additional models:

- Model 2- Select the important features and build model again.
- Model 3- Model with significant attributes given by Model2; analyze model performance.
- Model 4- Perform Grid search for all Machine learning models with significant attributes and get evaluation metrics for all models.

(b) Tabulate the train and validation results for all the above models.

(c) Analyze model performances and finalize the model.

Additional References and Practice:

https://www.kaggle.com/tags/logistic-regression
https://machinelearningmastery.com/logistic-regression-for-machine-learning
https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf
http://www.stefan-evert.de/SIGIL/sigil_R/materials/regression3.slides.pdf