Assignment - Neural Networks

Pages

Problem Statement

The assignment uses the same problem statement as the PA-I case study, though the **data is different** and hence you are required to download it from this page. This is an **individual assignment** and you will be **graded** on your performance. The deadline for submission of this assignment is **December 18, 2016, 11:59 PM**. For submissions obtained within 5 days of the deadline(December 24, 2016), there will be a 30% penalty. Submissions beyond 5 days of the deadline will not be accepted.

Company Information:

A telecom company called ‘Firm X’ is a leading telecommunications provider in the country. The company earns most of its revenue by providing internet services. Based on the past and current customer information, the company has maintained a database containing personal/demographic information, the services availed by a customer and the expense information related to each customer.

Problem Statement:

You are working for the telecom company ‘Firm X’. It has a customer base set across the country. In a city ‘Y’, which is a significant revenue base for the company, due to heavy marketing and promotion schemes by other companies, your company is losing customers i.e. the **customers are churning**. Whether a customer will churn or not will depend on data from the following three buckets:

1. Demographic Information
2. Services Availed by the customer
3. Overall Expenses

The data is provided at the end of the page**.** The aim is to automate the process of predicting if a customer would churn or not and to find the factors affecting the churn. The collated **data dictionary** for the variables in the 3 data frames is given below:

S.No.	Variable Name	Meaning
1.	CustomerID	The unique ID of each customer
2.	Gender	The gender of a person
3.	SeniorCitizen	Whether a customer can be classified as a senior citizen.
4.	Partner	If a customer is married/ in a live-in relationship.
5.	Dependents	If a customer has dependents (children/ retired parents)
6.	Tenure	The time for which a customer has been using the service.
7.	PhoneService	Whether a customer has a landline phone service along with the internet service.
8.	MultipleLines	Whether a customer has multiple lines of internet connectivity.
9.	InternetService	The type of internet services chosen by the customer.
10.	OnlineSecurity	Specifies if a customer has online security.
11.	OnlineBackup	Specifies if a customer has online backup.
12.	DeviceProtection	Specifies if a customer has opted for device protection.
13.	TechSupport	Whether a customer has opted for tech support of not.
14.	StreamingTV	Whether a customer has an option of TV streaming.
15.	StreamingMovies	Whether a customer has an option of Movie streaming.
16.	Contract	The type of contract a customer has chosen.
17.	PaperlessBilling	Whether a customer has opted for paperless billing.
18.	PaymentMethod	Specifies the method by which bills are paid.
19.	MonthlyCharges	Specifies the money paid by a customer each month.
20.	TotalCharges	The total money paid by the customer to the company.
21.	Churn	This is the target variable which specifies if a customer has churned or not.

How to start the assignment:

To solve any analytics problem, the Crisp-DM framework is to be followed.

The goal of this assignment:

You are required to develop multiple predictive models and find the best predictive model using the library "h2o". The data set provided below is only the **training data** while a part of the data is **intentionally not provided** to you since it will be used to **test the final model** you submit. You are expected to keep a part of the provided data as your own test data and use the rest as train + validation set.

Downloads:

You can download the datasets from below:

Telecom_Train
*file_download*Download

Note:

Please make sure the below points are to be followed strictly for evaluation purpose:

- The data set provided above is only the **training data**. The testing data, which is not provided to you, will be used to **evaluate the model you submit**.
- Store the collated dataset into “churn” object.

Packages Required:

You have to install the below packages below starting this assignment.

- install.packages(“MASS”)
- install.packages(“car”)
- install.packages("h2o")
- install.packages(“caret”)

Checkpoints

1. Business Understanding

Customer churn can depend on a lot of internal and external factors. External factors are the ones you might have no control or information about, for example, the launch of Reliance Jio will lead to churn across all telecom companies and predicting the churn for such cases gets extremely difficult. However, Internal factors such as the demographic information, the number of connections taken by a customer, personal information, billing information, information on services availed etc. can be used to predict the churn of customers. In this case study, you only need to consider the internal factors.

2. Data Understanding

The data is provided in the previous segment. The file customer.csv contains the personal information of the customers. The file churn_data.csv contains the information related to the churn of customers along with their billing information. The third file contains the information related to the internet usage and all the services customers are using.

3. What is the Company's Business Objective?

The company wants to understand the driving factors behind churn and wants to build a model which would predict future churn. The company can utilise this knowledge for churn prevention. Specifically, the company wants to determine which driver variables are having the most influence on the tendency of churning.

Problem Statement:

How do we approach the assignment?

The entire assignment is divided into 3 checkpoints to narrow down the problem statement. The checkpoints are interlinked with each other.

Please perform all the preliminary steps before model building. These preliminary steps do not carry any weightage.

- Load the data file.
- Make bar charts displaying the relationship between the target variable and various other features.
- Perform de-duplication of data.
- Bring the data in the correct format
- Find the variables having missing values and impute them.
- Perform outlier treatment if necessary

Checkpoint 1: Model Building

- **Model - Neural Networks - Tuning hyperparameters without epochs:**
  - Fine tune the hyperparameters without using epochs
    - Tip: The h20 library automatically splits the data into train and validation sets; you should use the performance on validation set as a guiding compass and change the hyperparameters accordingly
  - Experiment with various sets of hyperparameters i.e. number of hidden layers, activation function, number of neurons in each layer and drop out rate.
    - Report (as detailed comments in the R file) the symptoms each model displays and the actions you take on observing those symptoms

Note that the model **will be evaluated on test data** which has not been provided to you. You should use the validation set to guide you so that it does not overfit or be too simple either. The assignment will also be evaluated on the performance of your final model on the test set.

Also, make sure to **write detailed comments to explain your thought process** on building the models. If you reduce the number of layers because you suspect overfitting, write down why you suspect overfitting and why you decide to take the next step.

Checkpoint 2: Model Building

- **Model - Neural Networks - Tuning hyperparameters with epochs:**
  - Use epochs this time. Fine tune the hyperparameters and report the best model. Report the optimal number of epochs too.
  - Experiment with various sets of hyperparameters i.e. number of hidden layers, activation function, the number of neurons in each layer and drop out rate.
    - Report (as detailed comments in the R file) the symptoms each model displays and the actions you take on observing those symptoms

Checkpoint 3: Best Model

- **Model - Neural Networks - Best Model**
    - Report the best neural network model along with an explanation of why it is your best model. Report its performance metrics (accuracy, sensitivity and specificity).

Submission

This is an **individual assignment** and you will be **graded** on your performance. The deadline for submission of this assignment is **December 18, 2016, 11:59 PM**. For submissions obtained within 5 days of the deadline(December 24, 2016), there will be a 30% penalty. Submissions beyond 5 days of the deadline will not be accepted.

Important Note:

You are supposed to code entirely in R. Write the code in a well-commented R file which you can submit at the end. Please make sure to rename your R script with "**Roll_Number_main.R**".

Also, you have to write a brief explanation of your modelling results in the R file itself as comments, along with the best model.

The mark distribution is as follows:

Checkpoint/Details	**Marks Allotted**
Checkpoint 1 - Data Understanding and Preparation of Master File	100
Checkpoint 2 - EDA	100
Checkpoint 3 - Data Preparation	100
Checkpoint 4 - Model Building (300 marks per model)	1200

A broad **description of an ideal solution** is provided below to guide you:

Checkpoint/Details	Ideal Solution
Model building without epochs	Experiment with hyperparameters, observe results and iterate them based on sound reasoning. The choice of hyperparameters should be clearly explained based on the results obtained with each model.
Model building with epochs	Experiment with hyperparameters (including the number of epochs), observe results and iterate them based on sound reasoning. The choice of hyperparameters should be clearly explained based on the results obtained with each model. The choice of final model is based on sound reasoning based on these experiments.
Final Model	The final model performs reasonably well on the quarantined test data.

In total, you have to upload the one file as **one zip file named "Roll_no.zip".**The zip file should contain the **main R file.**

Submission*today*18th December 2016

Neural Network Assignment

This is an individual assignment. The deadline for submission of this assignment is December 18, 2016, 11:59 PM. For submissions obtained within 5 days of the deadline(December 24, 2016), there will be a 30% penalty. You are supposed to code entirely in R. For the algorithm, write the code in a well-commented R files which you can submit at the end. Please make sure to rename your R script with "Roll_Number_main.R".