**READ ME File - Nano Degree Program-Data Scientist- Project 1(Data Science Blog)**

Contents

[1. Installation 2](#_Toc21973793)

[2. Project Motivation 2](#_Toc21973794)

[3. File Descriptions 3](#_Toc21973795)

[4. How to Interact with Your Project 3](#_Toc21973796)

[5. Licensing, Authors, Acknowledgements 4](#_Toc21973797)

# Installation

No additional libraries are needed to run the code here beyond the Anaconda distribution of Python. The code uses Python 3.0

Below are the few steps to run the jupyter notebook.

1. Load the Python notebook into local Jupyter or Cloud ML platform(e.g. SageMaker)
2. Pls. make sure to load the data to appropriate directory
   1. Since I have used the Local Jupyter Notebook, I have uploaded the data into same folder as notebook.
   2. But you may need to modify the code, if you need to store data in a different place
      * *# read in the csv file* churn\_data = 'churn\_data\_v3.csv'
      * *# print out some data*
        + churn\_df = pd.read\_csv(churn\_data)
      * print('Data shape (rows, cols): ', churn\_df.shape) print() churn\_df.head()
3. Run the Notebook and analyze the result.

# Project Motivation

Currently a Company is facing a higher Customer Churn issue. The Churn rate is growing up in very quarter especially for Video subscribers.

The Customers are disconnecting their services due to multiple issues, related to Billing, Promotion drops, High price points, Technical issues , End of Contract period etc. This is impacting Company’s bottom line revenue, while operating cost remains static. This is causing enough pressure for Company to prevent revenue leakage by retaining the customers.

In this project, I will design and deploy a supervised Model(such as Logistics regression) to predict if a Customer will churn in future by training the Model with Customer data(Churned and Non-Churned)I have used the Churn Customer dataset from the Company database for approx. 10,000 accounts.

Key Business Questions:

1. What are the Key attributes/features of already-churned Customers?
2. Which Customers are most likely to churn in future?
3. What can be done to reduce the churn rate?

Detailed responses are in **Blog**

# File Descriptions

**Churn\_data\_v3.csv** contains Customer churn data. This data set contains key information, such as Account status, contract term, monthly charge amounts etc.

There is also a jupyter notebook available here to showcase all the work related to my three questions.

# How to Interact with Your Project

I have selected a Supervised ML “Binary Classification” model ,such as Logistic regression to predict the Customer Churn (i.e. Customer will churn or not in future).

Justification: Logistic regression extends the ideas of linear regression to the situation where the outcome variable, *Y* , is **categorical**. In our project , we need to solve a binary Classification problem, which can be solved using Logistic Regression Model.

It uses logs of odds as Target/Dependent variable. it predicts the probability of occurrence of a binary event utilizing a logit function

Metrics/Results:

We have used the K-fold Cross validation to measure the effectiveness of the model. This step randomly divides the set of observations into k groups. The 1st set is treated as Validation set and the method is fit(Training) on remaining k-1 fold

We have used the **cross\_val\_score** helper function of sklearn to determine the accuracy/effectiveness of the model. In this project, I have used CV = 5, i.e. 5- fold cross-validation. Accuracy was **84%**

We have also used the Precision and Recall metric to measure the performance/quality of the chosen Model

This metrics is primarily used for Classification models with especially an imbalanced data set(*i.e. # of not-churned datapoints are higher than # of already-churned datapoints)*

* **Recall (Sensitivity**): TPR (*True Positive Rate*) = This will be our key metric as we will be measuring our performance against, since our primary goal is to correctly predict the positive cases(i.e. Customer will churn)
  + Recall = TP / (TP+FN)= **88%**
* **ROC**(Receiver Operating Characteristic)
  + \*When we want to give equal weight to both classes Prediction ability ,we can look at the ROC curve. In our case(as stated in <Result Section> , AUC(Area Under Curve) is **88%**
* Challenges/Difficulty faced during Developing the Model:
  + Regarding the Data set, I think it would have been better to use Customer Call volume data as the part of predictors. But this data was bit hard to get. So I went ahead with data set that I have received.
  + There are quite challenges in cleaning up the data set and deriving the new Predictor variable(tenure) using other predictor variables(such as activation or disconnect dates. I had to write the code for that.

# Licensing, Authors, Acknowledgements

I would like to give credit to Udacity online courses and Kaggle for some of code ideas, and to Kaggle for the data. Few additional links/resources are stated below that I have used as reference to complete my project.

1. Choosing the right metric [Choosing the Right Metric for Evaluating Machine Learning Models](https://www.kdnuggets.com/2018/04/right-metric-evaluating-machine-learning-models-1.html).
2. Article about [What metrics should be used for evaluating a model on an imbalanced data set](https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba)
3. Article on [various techniques of the data exploration process](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/).
4. Churn model accuracy benchmark source ; <https://www.kaggle.com/>
5. Data Mining for Business Analytics- *Galit Shmueli, pertr C Bruce ,Wiley Publications*
6. DataCamp: Logistic regression :<https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>