INTERNSHALA TRAININGS

The Data Science Challenge

6th June 2019 - 24th July 2019

Background and Motivation

Data Science challenges are fun. They are full of learning & intense competition, and a lot is packed in a matter of a few days. You get to apply your learning on real-life datasets and also compare yourself against the other data science students from around the country to see where you stand. The thrill of finding a solution in a and competitive environment is addictive.
This data science challenge, in partnership with our data science training partner Analytics Vidhya, is our attempt to inspire the next generation of data scientists to fall in love with Data Science, to learn how to work with real-life data sets, and win exciting prizes.

Project Problem Statement

Your client is an Insurance company and they need your help in building a model to predict whether the policyholder (customer) will pay next premium on time or not.

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that you pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of Rs. 5000 each year for a medical insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalised in that year, the insurance provider company will bear the cost of hospitalisation etc. for upto Rs. 200,000. Now if you are wondering how can company bear such high hospitalisation cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a

few of them (say 2-3) would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is life insurance where every year you pay a premium of certain amount to insurance provider company so that in case of unfortunate event of your death, the insurance provider company will provide a compensation (called 'sum assured') to your immediate family. Similarly, there can be a variety of insurance products for different kinds of risks.

As you can imagine, if a large number of customers do not pay the premium on time, it might disrupt the cash flow and smooth operation for the company. A customer may stop making regular premium payments for a variety of reasons - some may forget, some may find it expensive and not worth the value, some may not have money to pay the premium etc.

Building a model to predict whether a customer would make the premium payment can be extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers who are less likely to pay and convince them to continue making timely payment.

Now, in order to predict, whether the customer would pay the next premium or not, you have information about past premium payment history for the policyholders along with their demographics (age, monthly income, area type) and sourcing channel etc.

For complete problem statement, datasets, data dictionary etc., please access the contest environment as described below.

The Contest Environment:

We have set up a complete contest environment for you on Analytics Vidhya, exclusively for this challenge. To access this contest environment, please follow below steps -

1. Go to https://datahack.analyticsvidhya.com/contest/internshala-data-science-challenge/

2. You will be required to sign up/create an account on Analytics Vidhya (if you don't have one already) **using the same email address with which you registered for this challenge on Internshala.** If you register with a different email, we will not be able to identify you in the contest environment and evaluate your submission.

3. After that, click on 'Register' button on the contest page to complete registration for Internshala Data Science Challenge and accept the T&Cs of the challenge
4. Once you are inside the contest environment, you can access the problem statement, datasets, data dictionary and other details relevant to the challenge.
5. Please submit your predictions using the solution checker provided on the challenge page.
6. You can use solution checker to make multiple submissions and see your position on Public Leaderboard before you make final submission
7. Final submission must be selected from all submissions and the code file required to create the submission must be uploaded along with the final predictions

How will the submissions be evaluated:

- Test data provided to you is further randomly divided into Public (40%) and Private (60%) data. You don't need to do anything for this. This is something that the contest environment would take care of. You need to submit your predictions for the entire test data.
- Leaderboard and scoring will be based on AUC-ROC Score
- Your initial responses will be checked and scored on the Public data.
- The final rankings and winners would be decided based on your private score which will be published once the competition is over and your code review.
- The top few contestants *may* also be interviewed (telephonic or in-person) and final list of winners will be declared post code review, and interviews on **14th August 2019.**

Rules of Data Science Challenge:

1. Your Data Science training began on 6th June 2019.
2. The contest project problem statement and instructions (i.e. this file) can be downloaded from the progress tracker once you unlock 4th (final) module of the training.
3. The last date to make your final submission (as per the instructions given above) is 24th July. Any late submissions will not be considered.
4. Only Individual participation is allowed. Use of external data is not allowed. Use of 'ID' variable as a predictor is not allowed.
5. You are expected to solve the project using only those algorithms and concepts that have been taught in the training. Use of any other algorithm(s) to make predictions is not allowed.
6. Throughout the challenge, you are expected to respect fellow participants and act with high integrity.

7. The winners will be decided on the basis of private leaderboard and code review post the contest closes on 24th July 2019.
8. The top few contestants *may* also be interviewed (telephonic or in-person) and final list of winners will be declared post code review, and interviews on **14th August 2019.**
9. Internshala & Analytics Vidhya hold the right to disqualify any participant at any stage of the competition if the participant(s) are deemed to be acting fraudulently.
10. The decision on the winners and runners-up made by Internshala & Analytics Vidhya will be final and binding.

Important Dates:

- Contest (and training) start date - 6th June 2019
- Contest end date - 24th July 2019. This is the last date to make your final submission.
- Result declaration - 14th August 2019

Extra Support for the contestants:

In case of any technical issue or issues related to the understanding of problem statement, please post your query using the 'forum' provided inside your Internshala Trainings environment.

Here are some additional resources you could refer:

- Data Science using Python
- Decision Trees & Tree based Modeling
- Essentials of Machine Learning Algorithms

All the best!