

DSSI Data Scientist Case Study

Background

The team has been asked to evaluate the recent revenue performance for a health insurance product. We are particularly interested in how long customers are keeping these insurance policies.

Data description

The data consists of policy details of customers. It includes:

- ID – Policy Id
- Policy Start Date – Date on which the policy started
- Policy End Date – Date on which the policy got cancelled/terminated
- Revenue – Total amount received from the policy
- Feature 1,2 and 3 – Attributes related to the policy

Requirements

- Write python code to answer the questions

Submission instructions

- Please place all the files used in your analysis in a zip file and return to your recruiting contact

DSSI Data Scientist Case Study

Questions to be answered

1. Calculate total revenue:
 - a) All 2021 policies (2021 Policy Start Date regardless of their end date)
 - b) 2021 policies by Policy Start Date
2. Calculate the First Year Renewal % of all policies that have a start date of 2020-07-01.
E.g., First Year Renewal % of 2019-07-01 start date can be calculated by:
Total 2019-07-01 policies = 7,834
Total 2019-07-01 that renewed (policies that have Policy End Date on or after 2020 July) = 4,271
First Year Renewal % = $4,271 / 7,834 = 54.52\%$
3. Describe in words what type of model would be appropriate to predict first year renewal? Explain the reason behind the model choice.
4. Assume that all policies with Policy End Date as November 2021 are active and the payments after November 2021 are unobserved. What type of statistical or machine learning modeling technique(s) would you use to build a predictive model of customer retention?
5. Assume that all policies with Policy End Date as November 2021 are active and the payments after November 2021 are unobserved. Predict the policy end date for all active policies.

Note: If you make any assumptions for any questions, please note them in the solution Jupyter Notebook