Trupanion Data Scientist Skill Assessment

Please complete the following 3 part homework to the best of your ability. Your submission will provide us with a work sample to evaluate as well as give you a feel for the type of work we do.

Part 1 – Preparing a usable dataset

Description:
When we receive invoice data from our customers, it is structured by line item. Trupanion then aggregates one or many line items to a higher level category for payment or denial. One of our end goals is to automate this line item grouping by categorizing each invoice line item. In order to do so, we must essentially undo that aggregation, so you must first map each line item to the appropriate group. Our labels (in file …labels) must be linked to our features (in file …lineitems).

Use the files named p1_labels.csv and p1_lineitems.csv to map as many line item data ids to labeled ids as possible.

Hints:

- "ClaimId" is a common index shared between the csvs
- The expected solution is a list of mappings from id_lineitem to id_labeled
- Not all line item data ids can be mapped
- Submit your work alongside your solution
    - We typically use python/jupyter notebook, but any similar language or tool is acceptable
- I've included a "features" column as a reminder of our end goal – nothing needs to be done with it

Part 2 – Build a text classifier

Description:
Our product does not cover routine, wellness or preventive care. We believe that costs that pet owners can expect periodically and budget for should be separate from an insurance policy meant to cover accidents and illnesses.

Use the data contained in p2_data.csv to build a binary classifier to predict the "PreventiveFlag" label using the text features provided. This model can be used to automate the detection of ineligible line items. The expected output are prediction probabilities for rows 10001 through 11000, where the labels are currently null.

Part 3 – Short answer

As an insurance company, we collect premium on a monthly basis from customers. In return, we pay the customer's veterinary bills should their pets receive medical treatment.
How would you estimate the expected value of a customer at any given point in time?
In what ways could we utilize this estimate?