Name: Cushing, Bradley
Date: 11/04/24
NYU ID: N10695516
Course Section Number: CSCI-GA-2433-001
Project #2 Report

Note: I search quite extensively for good external data to use to train a Logistic Regression model but I had a very hard time finding something that would be good for my use case. Therefore, after discussing with the professor I chose to create my own dataset that could be used for training such a model. I hand picked the attributes that I thought could be interesting indicators of whether a person who would purchase a warranty would be “risky”.

Semi-structured data sets

Risk categorization based on living situation

- **cushing_p2_su24_warranties_actual.csv**
- prediction/warranties-actual.csv (copy for training)
- prediction/warranties-random.csv (random testing)

There are a few copies of the external data used to train the prediction model using logistic regression. Please see cushing_p2_su24_warranties_actual.csv for the actual data.

Example CSV data

Age	Kids_count	Pets_count	Siblings_count	Income	Has_risk
16	3	3	0	359686	1
88	4	3	3	75054	1
19	4	2	3	169817	0

How insights inform EDA created

The first 4 columns of data are used as an indicator for whether the person purchasing a warranty for the product should be considered as “risky” or not. We will process this information submitted by a user who has purchased the product if they choose to also purchase a warranty. The base price of a warranty will be set at $5 dollars for 1 year for those that aren’t risky. For those that are risky we will double the price to $10 dollars.

Semi-structured files

We consider unstructured data to be semi-structured in reality as what we want to process from external sources will typically be in CSV, XML, or JSON which we can read, parse, normalize, and reformat into something structured like a table in our relational database. Then we can

query our relational database with this structured data to gain insights and inform the business. We won't consider completely unstructured data like text documents, audio files, etc.

Prediction folder

Training is done in the prediction folder in the actual file is called "warranties-actual.csv" which is used to train the model. I've included this file which has over 1000 fields along with the report. The fields we use for prediction are Age, Kids_count, Pets_count, Siblings_count, and Income amount. These are all deemed predictors for whether or not the person is "risky" which is someone we consider is more likely to use their warranty and request a replacement for their product. The data itself has been sorted by the Has_risk field so it's easier to read. Note that the intuition is that those with more kids and pets will have higher risk and that those with more siblings and higher income will have lower risk.

Training

I've used two files to test training of the model, with both warranties-actual.csv and warranties.random. The files are self described in the name, indicating data that has an actual pattern and can be understood versus data which is generated at random. When trained on the actual data there is 90% accuracy of prediction on actual data. When trained on random data there is 48% accuracy which is to be expected. Training on the real data with distinct patterns that exist in the real world allows us to benefit from making this prediction and charging more.

Logical database schema

The logical database schema has been created in a file cushing_p1_su24_logical_schema.sql. This includes all the necessary SQL code (DDL) to create the database according to the relational schema specified in the first step.