DATA ANALYTICS

Auto Insurance Case Study

The Background

You are a data analyst working as a member of the advanced analytics team at an insurance company. Your boss has come to you with a customer segmentation task. The company realizes that it does not have categories to label their customers and wants the advanced analytics team to create customer profiles. The thinking is that once customer profiles are created, the marketing team can then find new customers based on our team's results and market to them appropriately. Currently, the company charges the same for each policy premium, but the company wants to move towards a more dynamic pricing approach where risky drivers pay more, and less risky drivers pay less. This is an important task, and this order has come directly from the Vice President of Sales and Marketing who will be participating in your presentation.

The Data

Along with this document, you should have received one dataset:

1. auto_policies_2020.csv

This data set is a set of personal auto insurance policies taken out in 2020. There are 60,392 policies (rows), of which 10,030 had at least one claim. There is also a substantial number of missing values.

The Case Study

Your manager has given you the following task:

1. Create a customer segmentation campaign that uses the auto_policies_2020.csv dataset to label customers based on their risk profiles (e.g. riskier customers, less risky customers, normal drivers, etc.).

Sharing Your Results

1. Please create a PowerPoint presentation that explains your results:
   - The first half should be targeted to V.P. of Sales and Marketing and her team. Please give them your recommendations and explain how and why you are making those recommendations. Keep in mind that the V.P. of Sales and Marketing and her team do not have a quantitative background and are only vaguely familiar with data science. Additionally, not all members of the marketing team are familiar with the problem statement and data. Be sure to review the problem statement and data at the beginning of the presentation.
   - The second half should be targeted to a group of data analysts and data scientists. Please give your detailed methodology and thought process.

2. Upload your presentation, code, and any other documentation you have Canvas before the deadline.

DATA ANALYTICS

Rubric

Your grade will be determined by two parts. Below are questions to consider. Note that this list does not include every possible question, and that you should explore deeper.

1. Code
2. Presentation containing marketing campaign

Code

- Do you have good coding practices?
- Did you comment your code for easy reading?
- Did you pre-process the data correctly?
- Did you use at least 2 different clustering algorithms?

Presentation

- Did you target your presentation to the appropriate audience?
- Did you identify target profile groups?
- Did you provide one final recommendation to management and marketing?
- Did you explain at a level appropriate for your audience?
- Did you include appropriate visualizations in your presentation?
- Did you provide a detailed summary of why your methodology choice to the technical users?
- Did you motivate your model selection?
- Why did you remove/select some features versus others?
- Why did you choose these clustering algorithms?
- How did you choose the parameters you used for your algorithms?
- How did you determine the best parameters?
- Did you try any approaches to clustering categorical data?

Did you submit your completed Midterm documents to Canvas before the deadline?

DATA ANALYTICS

A glossary for the datasets is provided below.

pol_number	policy number for the insurance policy
pol_eff_dt	auto insurance policy effective date
gender	gender of driver: F, M
agecat	driver's age category: 1 (youngest), 2, 3, 4, 5, 6
date_of_birth	driver's date of birth
credit_score	driver’s credit score(integer): 1-100, 1=poor, 100=excellent
area	driver's area of residence: A, B, C, D, E, F
traffic_index	traffic index of driver’s area of residence(integer): 100=country average, >100 means worse traffic conditions than average
veh_age	age of vehicle(categorical): 1 (youngest), 2, 3, 4
veh_body	vehicle body, coded as: BUS CONVT = convertible COUPE HBACK = hatchback HDTOP = hardtop MCARA = motorized caravan MIBUS = minibus PANVN = panel van RDSTR = roadster

DATA ANALYTICS

	STNWG = station wagon TRUCK UTE = utility
veh_value	vehicle value, in $10,000s
months_insured	number of months vehicle insurance is bought(integer)
claim_office	office location of claim handling agent: A, B, C, D
numclaims	number of claims(integer): 0 if no claim
claimcst0	claim amount: 0 if no claim
annual_premium	total charged premium i.e. the cost of insurance