DATA ANALYTICS

Auto Insurance Case Study

The Background

The Sales & Marketing Team was pleased with your work in creating categories for existing customers. They now have confidence in your work and are trusting of your results. Your boss has now come to you with a supervised learning task. The company wants to engage future customers to increase our client base. Historically, the Sales & Marketing Team mails promotional flyers and call potential customers directly. However, the company would like to save money on material costs and instead wants to predict customers which customers to target. In particular, the marketing team has acquired data about potential customers and wants to predict the number of claims per customer and the claims cost per customer. The motivation behind this is that if we can accurately predict the number of claims and cost of the claims for future customers, we can direct our marketing to customers with fewer claims.

From a business perspective, the worst outcome is predicting that a potential customer will have no claims, but in actuality will have any amount of costly claims.

This is an important task, and this order has come directly from the Vice President of Sales & Marketing who will be participating in your final presentation. The expectations are higher for this presentation given that the representative groups are aware of your capabilities. Additionally, one concern about the previous presentation was how the missing values were handled - the VP of Sales & Marketing did not like this approach. The VP believes that the data is important and contains a lot of vital information. For this analysis, the VP is okay with removing some missing values, but wants you to remove fewer data observations than last time.

The Data

Along with this document, you should have received two datasets:

1. auto_policies_2020.csv (training)

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.

2. auto_potential_customers_2022.csv (testing)

This data set is a set of personal auto insurance quotes for insurance premiums starting in 2022. There are 7,464 quotes (rows). There is also a substantial number of missing values.

DATA ANALYTICS

The Case Study

Your manager has given you the following task:

1. Target the potential customers in 2022 which will result in the lowest claim amounts. That is,
    - Predict the number of claims for each potential customer, and
    - Predict the cost per claim, given that a claim occurs

Sharing Your Results

1. Please create a PowerPoint presentation that explains your results:
    - The first half should be targeted to V.P. of Sales & 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. Provide a detailed summary of the approaches you attempted even if they failed.

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 attempt multiple strategies?

Presentation

- Did you target your presentation to the appropriate audience?
- Did you provide one final recommendation to management and marketing?
- Did you provide the business value of your predictive model?
- Did you explain at a level appropriate for your audience?
- Did you include appropriate visualizations in your presentation created using only Python or Tableau?
- Did you provide a detailed summary of why your methodology choice to the technical users?
- Did you motivate your model selection?
- Did you explain how you approximated your error?
- Did you explain which metric you chose and why?
- Why did you remove/select some features versus others?
- How did you choose the parameters you used for your algorithms?
- How did you determine the best parameters?
- How did you handle missing values?
- How did you handle class imbalance?

DATA ANALYTICS

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

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