Case Study 2

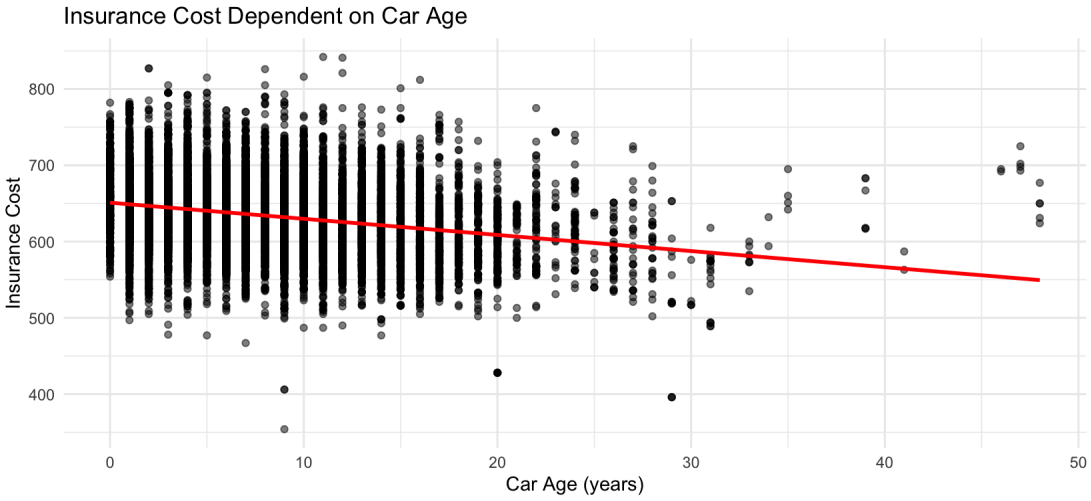
Section A, Team #12

Ankit Chandekar, Jinghua He, Abigail Miller
Muhammad Memon, Nikki Reddy, Katherine Zhang

1. Pick two (or more) variables and attempt to show a relation between them via visualization. As discussed before, this requires one to formulate a question, and to communicate clearly a conclusion based on data visualization (specify the why, what, how).

Typically, car insurance companies dictate the price of car insurance premiums based on the age of the insured's car. New cars are more expensive to insure because they have higher repair costs associated with them. On the other hand, older cars are cheaper to repair in cases of damage. Visualizing the relationship between car age and insurance cost helps us understand whether the data matches the industry expectations.

Below is a scatterplot with car age measured in years on the x-axis and insurance cost measured in dollars on the y-axis. The red line is the line of best fit that demonstrates the relationship between the two variables. As the car age increases, the insurance cost decreases.



2. Provide a model based on linear regression to forecast the quoting procedure from ALLSTATE based on the observed variables. Pick two variables of your model, describe their marginal impact on the quote, and comment the interpretation from the business perspective.

The formula I chose for my linear regression model includes all variables and interactions between coverage options. Model: result_interactions <- glm(cost ~ . +(A+B+C+D+E+F+G)^2, data = DATA)

One variable in the model is group_size, which indicates how many people will be covered under the insurance policy. The p-value of this variable in the model is .041, so it is significant. This variable has a coefficient of 2.611, meaning that as there is an increase in group size by one member, the insurance quote will increase by $2.61. This is because every additional person added to a car insurance policy increases the risk that someone in the group will get into an accident, increasing the likelihood that the group will create costs for the insurance company.

Another variable is homeowner1, which is significant with a p-value of 2e-16. The coefficient is -13.18, meaning that if someone is a home owner, their insurance quote decreases by $13.18. Home owners are often considered more responsible than non-home owners, indicating that home owners may have less risky driving patterns, reducing the number of accidents they are in. Also, home owners are usually more financially responsible than non-home owners, so they can pay for more minor accidents out of their own pockets instead of filing claims with the insurance company. Overall, home owners pay less because they are not high risk nor claim prone customers, reducing the expected costs that insurance companies expect to cover.

3. Suppose that a customer will pick the lowest between the quote you provide and that ALLSTATE provides. Build a model framework (follow/adapt steps in Model Framework in Class 3 for the Churn Problem) to maximize expected revenue1 from a customer given the observed characteristics. This includes the mathematical model, description of a decomposition strategy, the associated core tasks, and specific data mining methods you would choose. For each core task comment if it can and if it cannot be implemented with the available data.

First, we need to understand the business. The goal is to choose our quote to maximize the expected immediate revenue when the customer always buys the lower of our quote and ALLSTATE's.

1. Framing:
    - Action: choose our quote p for a customer with features x
    - Event: we win if $p \leq A$ and A is ALLSTATE's quote
    - Pay off: If win$\rightarrow$ revenue=p, if lose $\rightarrow$ 0
2. Modeling choice: Quantile regression for the competitor's price
    - We estimate the conditional quantile function using Quantile Regression
    - Why:
        - Quantile regression directly models $Q_{A|x}(\tau)$ for many τ.
        - Interpolating across quantiles lets us recover a CDF $F(\cdot|x)$ without Normality.
        - It naturally handles heteroskedasticity and tail behavior.
3. Decomposition strategy
    - Predictive distribution via quantiles:
      Learn a grid of conditional quantiles $\{q_\tau(x):\tau \in T\}$ for $A|x$ using Quantile Regression.
    - Decision rule:
      Evaluate $ER_\tau(x)=q_\tau(x)(1-\tau)$ over $\tau \in T$, apply guardrails (price floor, monotone coverage constraints)
4. Core task, chosen methods and implementability
- Data preparation & assumptions (could be implement)
    - Factorize categoricals
    - Exclude customerID, time location
5. Predictive task(quantile regression) (could be implement)
    - Model from R, rq(cost ~ . +(A+B+C+D+E+F+G)^2, tau = grid, where grid = e.g., {0.05, 0.10, …, 0.95}. Main effects for all observed variables + pairwise interactions among A–G

6. Probability of winning (could be implement)
    - At price $p=q\tau(x)$, set $Pr^{\wedge}(win|x,p) \approx 1-\tau$.

4. Suppose that a customer will pick the lowest between the quote you provide and that ALLSTATE provides. Aiming to maximize expected revenue2, provide quotes for each of the three customers specified in "new.customers". Clearly state which core task and which data mining method you used to provide the quote.

Core task: Our task is to predict ALLSTATE's quote for each new customer and set our quote to maximize immediate expected revenue when the customer buys the lower of the two quotes we provided.

Data mining method used: We trained a quantile regression model with all observed variables plus pairwise interactions among coverage options (A-G) with rq(cost ~ . +(A+B+C+D+E+F+G)^2, tau = grid where grid is a set of quantiles. Before modeling, we applied the assumptions and preprocessing, like treating risk_factor / C_previous NA as a level "0", making factors for car_value, day, state, etc. and ignoring customer_ID, shopping_pt, record_type, time, and location so that the estimation does NOT differentiate by ID/visit/time/location.

Decision rule/pricing for quote: For each new customer and each quantile τ, let $q\tau$ be the predicted ALLSTATE quote at that quantile. The probability we win at that price is approximately 1 - τ. We would compute the expected revenue: $ER(\tau) = q\tau \times (1-\tau)$, apply a small price floor, and choose the quantile τ that maximizes ER. Our final quote is $q^* = q\tau^*$.

Using quantile regression, we predicted a grid of conditional quantiles for ALLSTATE's price per new customer and evaluated $q\tau \cdot (1-\tau)$ across τ. The table below shows the median prediction, the chosen quantile, and our recommended quote. In our data, the expected revenue maximum occurred at $\tau = 0.10$, which is more aggressive pricing. We do not want to have a lower τ because then our revenue per customer will be extremely low. If we take a higher τ, then we will have a profit margin per policy but may not acquire as many customers, also limiting expected revenue.

	Customer	Median	ChosenQuantile	FinalQuote	ExpectedRevenue
1	1	625.84	0.1	583.27	524.94
2	2	635.56	0.1	611.80	550.62
3	3	636.49	0.1	605.81	545.23

5. Suppose next that the customer might not accept either of the two quotes (but he will consider only the smallest of the quotes). Build a model framework (follow/adapt steps in Model Framework in Class 3 for the Churn Problem) to maximize expected profit from a customer given the observed characteristics. This includes the mathematical model, description of a decomposition strategy, the associated core tasks, and specific data mining methods you would choose. For each core task comment if it can and if it cannot be implemented with the available data.

1. Mathematical Model

The core idea of the model is to figure out the best price to offer a customer to make the most money. The profit depends on two things:

1. **The profit we make if the customer buys from us:** This is Quote minus the E[Cost] (the price we charge them minus the costs if they make a claim).
2. **The chance that the customer will actually buy from us:** This is the P(Accept) (the probability of acceptance).

The challenge is that these two things work against each other. If we offer a very low price, the customer is more likely to accept, but the profit per sale will be small. If we offer a very high price, our profit per sale will be large, but the customer is unlikely to accept, especially if our competitor has a lower quote.This model helps find the price that balances a lower profit per sale with a higher chance of making the sale, ultimately leading to the highest possible overall profit.

2. Decomposition Strategy

Our main problem can be decomposed into two distinct core tasks: one focused on the customer's behavior(classification) and the other on the cost of a claim (regression). The results of these models are then integrated into the model to calculate the expected profit.

- The goal of the classification model is to predict the probability that a customer will choose our quote over our competitor's, given various factors.
- The goal of the regression model is to predict the expected cost of a claim for a specific customer, which represents our cost of goods sold.

3. Data Mining Methods

Core Task 1: Predicting Customer Acceptance

- **Goal**: Predict the chance a customer chooses our quote.
- **Methods**: Classification models like Logistic Regression or Random Forests.
- **Data Problem**: We cannot complete this task. The data lacks a key piece of information: whether the customer actually accepted a quote. It only shows what quotes a customer received, not what they bought.

Core Task 2: Predicting Expected Claim Cost

- **Goal**: Predict the cost of a claim.
- **Methods**: Regression models like Multiple Linear Regression or Generalized Linear Models.

Data Problem: Cannot be implemented: The primary data constraint is the lack of a true "claim cost" or "loss" variable. The cost variable in the data set is the quoted price, not the actual cost of a claim incurred by the insurance company.