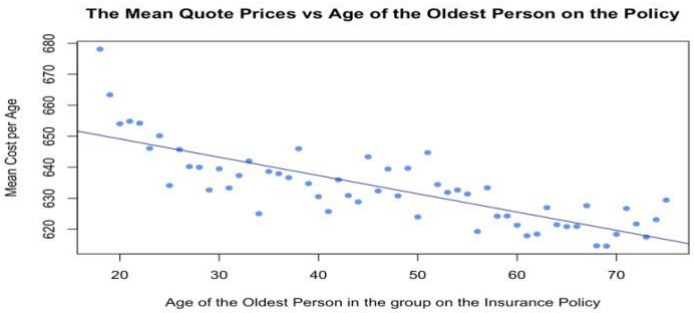
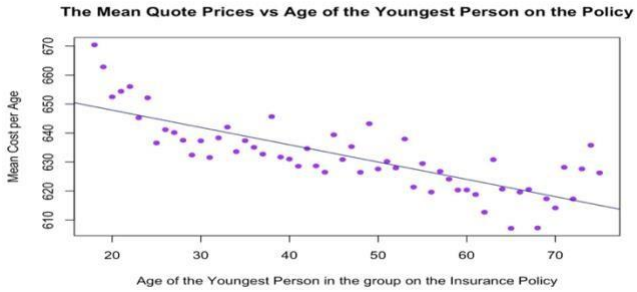
Insurance-Undercutting Report

We were keen on understanding the relationship between the cost of the insurance vs the age of the person. In order to get a better idea of how the age impacts the quote of the insurance, we made a graphical plot consisting of the cost Vs age_oldest and age_youngest. On the x-axis we plot the variable "age_oldest" and on the y-axis we plotted the average "cost". We noted that there is a downward trend in this visualization which shows that as the age of a person increases, the cost decreases. This result was in line with the case, which explained that a "younger person" would live a "riskier life". The same trend can be observed when we plotted the cost Vs age_youngest.





Our equation is as follows:

Cost of insurance quote = 677.6 + 14.2homeowner - 1.1car_age + 29.9car_valuea - 50.6 car_valueb - 44.4 car_valuec - 39.4 car_valued - 40 car_valuee - 38.4 car_valuef - 34.2 car_valueg - 26.6 car_valueh + 1.6 car_valuei - 17 risk_factor1 - 6.8 risk_factor2 - 0.6 risk_factor3 - 0.17 risk_factor4 + 0.75 age_oldest - 1 age_youngest - 6 married_couple - 4.7 C_previous1 - 15.4 C_previous2 - 18.1 C_previous3 - 23.3 C_previous4 - 1.1 duration_previous + 47.3A - 9.8B + 4.3C + 8.4D + 27.7E + 14.6F - 6.2G - 1.7A*B + 0.5A*C - 5.7A*D - 6.6A*E - 10.3A*F - 1.5A*G + 1.9B*C + 0.7B*D - 1.8B*E + 1.3B*F + 1.8B*G - 2.6C*D - 1.6C*E - 2.9C*F + 2.5C*G + 2D*E - 2.6D*F + 0.7D*G - 1.5E*F – 1E*G + 1.8F*G

For considering the impact of a variable on the dependent variable that we are trying to predict, we must take a look at the value of the coefficients. The value of cost will increase or decrease based on the significance and impact of each independent variable.

The variable "married_couple" has a coefficient of -6.04 in our regression equation. This indicates that by being a married couple (i.e, the variable is equal to 1), the cost of the insurance policy would reduce by USD 6.04. The interpretation of this from a business perspective would be that if a vehicle owner is married it would indicate that they are more stable, more mature in age and are at a lower risk of actually needing the insurance due to relatively careful driving (as opposed to someone who is 19 years old).

On the other hand, "home_owner" has a coefficient of -14.16 in our regression equation. The impact of having a home i.e. having the value "1" in the column "home_owner" would have a negative effect on the price of the quote. This means that if a person owns a home, they have assets to their name, and that implies more stability and a lower risk for an insurance company. So the value of cost per change of 1 in "home_owner" is -14.16 dollars.

Business Understanding :

Our goal is to maximize the revenue of our insurance company and price our policies at a rate that is attractive to the customers.

The mathematical equation:

E [Expected revenue from customer] = P(customer leaving| Quote Price our contract) X (Quote Price our contract -difference – 0.5) + P(customer staying | Quote Price our contract) X (Quote Price our contract + difference – 0.5)

Where :
Difference = ALLSTATE contract price – our contract price P
= probability

** P(customer leaving|Price our contract) and P(customer staying | Price our contract) can only take 2 values which are either 0 or 1. This is because the problem statement explicitly states that the customer will only gravitate towards the cheaper contract. Therefore, when the price of our contract is higher than ALLSTATE, the customer will go ahead with ALLSTATE and vice versa.

Deployment:
In order to maximize revenue, we have decided to price our insurance contracts 50 cents lower than ALLSTATE's insurance contracts. This will ensure that at any given time, all the customers will choose our insurance contract over ALLSTATE's contract because it has cheaper and more competitive pricing.

Data Understanding

Higher Level Limitation:
As the case suggests, the customer picks the policy with the lower price. However, we understand that only focusing on price is a limitation in the real world, because the customer might look at more parameters than just the price of the policy such as wider coverage options, customer service etc.

Limitation of Our Model:
Since the data is limited, we must make a few assumptions. In our model framework, we have considered adjusting our quote prices to be 50 cents less than the competitors. Our assumption is that the customer would pick our quote even though it's only 50 cents lower than the competitor. We understand that in the real world, outside of the case, it would take a higher discounted price to attract a customer.

Investing in Data:
In an industry like Insurance, we would face high competition when it comes to pricing. A lot of our pricing estimates are done based on competitor data in order to increase our customer base. Hence, we would say that there is significant advantage in investing in data. The only exception here is that the data is not at an unreasonable price and that the company still needs a good pricing model/strategy so it makes sense to invest.

Modeling
Using the model stated in Q2, we have forecasted the prices of our insurance contracts. Based on our analysis, our insurance contracts are priced cheaper only 7262/15483=46.9% of the time in comparison to ALLSTATE. This means that based on our sample, the customer picks our contract over ALLSTATE only 46.9% of the time. We have taken the steps below in order to ensure that the customer chooses our contract 100% of the time (to maximize revenue).

Methodology in pricing/ explanation of the model:

a. We have done a side-by-side comparison of ALLSTATE's insurance contract prices Vs our forecasted insurance contract prices (Using the model in question 2) and identified separately, our contracts that are priced higher and lower in comparison to ALLSTATE.
b. For our contracts that are priced higher than ALLSTATE, we have lowered the price down to the price of the ALLSTATE contract (I.e, given the customer a "discount") and further reduced 50 cents from it to ensure that we're always priced lower than ALLSTATE.

c. For our contracts that are priced lower than ALLSTATE, we have increased the price up to 50 cents lower than the ALLSTATE contract in order to maximize revenue. This will ensure that we have no contracts that are priced unnecessarily lower than it needs to be. This will help us maximize revenue

The above steps will ensure that all customers will consider our contract over the ALLSTATE contract, given that the only factor they consider in deciding is the price of the contract (i.e, the contract that is priced lower). This will ensure that revenue is maximized. Further, the increasing of the price of the contracts that are cheaper than ALLSTATE while ensuring that we're still effectively priced lower than them as well will ensure that no contracts are priced at a large unnecessary discount.

The exact price points and the methodology can be found on the excel sheet we have submitted (Sheet name = DataScience_Team3A, excel sheet = Q3)

Assumptions:

We have assumed the below assumptions when building the model:

a. Customers would be inclined to go ahead with our insurance policy Vs ALLSTATE even though it's only 50 cents cheaper
b. Pricing is the only aspect a potential customer considers when choosing an insurance provider
c. Our only competitor is ALLSTATE

Decomposition techniques:

We have split our forecasted values into two subcategories in order to get a more accurate pricing for our contracts.

a. Our forecasted values that are higher than the ALLSTATE values
b. Our forecasted values that are lower than the ALLSTATE values

We have put together the above two values in order to estimate our final contract prices. **Our core task is Regression**. In this way, we only consider the significant variables in our model. We have used a generalized linear regression model in order to determine the expected quote prices. Our task in this case is predicting values and hence we must go with regression rather than other tasks like classification as we need the forecast data for the quote prices. Our data mining method is running regression model on original variables and variables generated using interaction and only retaining those that have p-values less than 0.05. We also use our framework to improve the quote prices.

DATA.cost	forecast2 (Our forecast)	How many times do people choose us	Difference b/w the competitor and us	Rounding	Our New Price	Do we get the customer?
623.4	608.8	1	14.6	15	622.9	1
639.6	656.5	0	-16.9	-17	639.1	1
638.8	642.4	0	-3.6	-4	638.3	1

Core task: Regression

Data mining method: Run regression model on original variables and variables generated using interaction and only retain those that have p-values less than 0.05. We have also further modified the quote price as per our framework. Using our regression model derived from the dataset, we calculate our expected quotes for these three customers. Comparing the results with competitors' quotes, we only get one result out of three that is lower. In order to maximize

the revenue, we make some adjustments for each quote so that they are all 50 cents below our competitors' targets, achieving the goal of obtaining all three customers for revenue maximization.

Business Understanding

Our goal is to maximize the profit of our insurance company by ensuring that we attract the correct customers by pricing our products effectively (i.e, price our contracts higher for customers with higher risk categories such as 3 and 4, and price our contracts lower for customers in risk categories 1 and 2 to maximize profitability). Our estimate is that customers in higher risk categories will be more expensive to the firm (more claims) and customers in lower risk categories will have less expenses to the firm (less claims).

The mathematical equation:

For each customer with features X (considering features like risk level), E [Profit | X] = P(Customer Staying| Price$_{\text{our contract}}$) X (Quote Price $_{\text{our contract}}$ - Expense $_{\text{cost to company)}}$ + P(Customer Leaves| Price$_{\text{our contract}}$) * 0

***Note: in our model, a customer who leaves brings no value (hence value=0) and has no cost associated. We have also considered that the Expense/ how much the person claims/cost to company does not consider other costs like marketing/target costs.*

Deployment/explanation of the model:

In order to maximize our profitability, we have decided to price our insurance contracts strategically as below by subcategorizing our potential customers in to 2 brackets:

a. Customers in higher (3 and 4) risk categories: Price the insurance contract 50 cents higher than the most expensive insurance contract in the market. These contracts are priced higher because we're willing to loose these customers due to them causing the firm a higher amount of expenses (i.e, more claims).
b. Customers in lower risk categories (1 and 2): Price the insurance contracts 50 cents lower than the most cheapest insurance contract in the market. These are the customers we believe would maximize the profit of our firm.

With the above prices, we anticipate that this will attract less riskier customers to our portfolio and deviate the higher risk ones.

Data Understanding

Higher Level Limitation:

We do not have access to the expenses (both fixed and variable) of the firm. In order to estimate which customers will cause less claims, we have assumed that customers with higher risk categories will be more expensive. However, in reality this may not be applicable. Further, similar to our framework in Q3, we believe that there are numerous factors beyond the price which customers consider when choosing an insurance provider.

Limitation of Our Model:

We have made the assumption that by offering higher risk customers the highest priced contracts in the market, they will deviate towards competitors. However, this may not be the case as customers might want to be a policyholder in the firm regardless of the higher quoted price.

Investing in Data:

We believe an investment is required in order to correctly determine how to price our insurance contracts in the market. Quoted prices of other competitors are required in order to make a more accurate estimate.

Modeling

Assumptions:

a. We have decided to omit the level risk category, as this is simply a substitute for “NA”. “NA” is used when the data does not exist.
b. We have assumed that the variable “risk category” takes into account all the data that makes a customer riskier/less riskier such as age of the oldest, age of the youngest, homeowner and age of the car etc.
c. The risk category of the customer directly affects the expenses of our company (i.e, directly correlated to the claims – riskier customers will have higher claims).
d. All expenses stated on our mathematical model are dependent on the customer (they’re all variable costs)
e. Customers in the risk categories 1 and 2 will have the same level of costs and they will continue to remain in those risk categories for the rest of their tenure with us. The same applies to customers in category 3&4.
f. We are not targeting any customers in terms of advertising.

Decomposition techniques:

In order to estimate which customers we needed to raise the price for, we divided the customers in to 2 subcategories based on their risk profile. That is bucket the cutomers as follows: Bucket 1: risk category 1&2, and Bucket 2: risk category 3&4. We have put together the above two scenarios in order to estimate our final contract prices. **Our core task is Regression,** which we have completed in Q3. We are taking into account that the people that are in Risk Category 1&2 are important to keep as they are low risk. Hence as the decision tree suggests, we reduce the price in this category to attract these customers. On the other hand, we assume that the customers with Risk Factors 3&4 have a higher likelihood of claiming their insurance and a higher cost/expense to the company. Hence, these are the customers that we are prepared to lose as if they bring in a loss, we cannot maximize our profit. This is our strategy to maximize profits.