2022 Travelers Business Insights & Analytics LDP Case Competition

Submitted by Team Time Travelers

Ajith Menon, Adwaith

Gamalapati, Sai Jahnavi

Krishnan, Nandita

Table of Contents

Executive Summary

In this case competition, we intend to analyze the dynamics that differentiate high-value customers. These dynamics will be of interest for the blue buffalo insurance company as they contribute to 80% of revenue.

In the dataset, we have 22 attributes which include personal information of record (like gender, age) and also the payment information of the customers (like premium, claims, fraud claims, etc.)

Coming to pre-processing, the dataset is merged into a single file. Next, we proceeded with basic preprocessing procedures like checking for null values or outliers. Further we came up with OUR ("time Travelers") definition of HIGH VALUE CUSTOMER. In addition to that, new variables were created by feature engineering methods to reduce the complexity of the dataset, for example making it effective.

In pre-processing, the two datasets are combined into a single file. Next, we proceeded with basic pre-processing procedures like checking for null values or outliers.

In our next step, we came up with our ("Time Travelers") definition of a HIGH-VALUE customer. Further, new variables are created by feature engineering methods to reduce the complexity of the dataset, for example making effective use of acceleration values. We also discarded some values that retain redundant information.

Finally, we found traits that can be attributed to high-value customers. Through this analysis, we found customers who could be classified as "high value customers". Based on the findings we came up with some recommendations for the buffalo insurance company.

Problem Statement

We are a group of Data Analysts working for Blue Buffalo Insurance, a property casualty insurance company based in downtown Hartford, Connecticut. Our business partner wants us to conduct a thorough analysis using historical policy and claim data. As the insurance market changes, the business partner is concerned about the company's retention rates. Customers are more willing to change the status quo than ever before. With current inflation and a volatile market, everyone is looking for ways to save a few dollars. Our team is in charge of analyzing and exploring historical data to determine which customers are "high value" to the company and how the company might be able to offer additional savings or incentives to keep these customers from switching insurance companies. For this analysis, we must incorporate the customers' driving telematics, which tracks the customers' driving habits. The goal here is to identify a subset of 'high value' customers and recommend retention strategies to Blue Buffalo Insurance executives.

Data Pre-Processing

policy_nbr (Categorical):

The above predictor is not useful for analysis. Hence this has not been considered for our analysis.

Age (continuous):

The typical age of the records is ranging from 18 to 64 with mean and median values as 40.4, 40.5

Gender (Categorical)

We observed there are more females (535) compared to males (465).

state_cd, zip_cd:

The above predictors are redundant with state_nm (state name). Hence these are dropped from our analysis.

state_nm (Categorical):

We are using state_nm for visualization purposes to understand and visualize the demographic distribution of attributes.

policy_effective_date (Categorical)

Policy effective date is when the policy is either expired or renewed. All the policies are renewed or expired in the year 2022 between months ranging from January to November.

We used this attribute to compute the number of years the customer is active which will be discussed in detail in the feature re-engineering section. For using this as datatype we changed data type from object to “DATETIME” in python

policy_start_date (Categorical)

Policy start date is when the policy is either expired or renewed. All the policies are in between 1977 to 2021 with the median year of 2007.

We used this attribute to compute the number of years the customer is active which will be discussed in detail in the feature re-engineering section. For using this as datatype we changed data type from object to “DATETIME” in python

auto_make, auto_model (Categorical)

- Auto make is the brand of the car. It has around 14 companies. Auto model is the variant of the car for a car brand. The attribute has around 39 distinct models.

The above feature combination is used for Feature Engineering the "TYPE" of the car.

auto_year (Categorical)

Auto year is the manufacturing year of the car. In the data the values are ranging from 1995 to 2015 with the majority of cars manufactured in 2005.

claims (Categorical)

Claims are the number of times the customer has claimed the insurance. We understand that lower the claims made by a customer will benefit the buffalo insurance company.

The typical range of claims in the dataset ranged from 0 to 6 with records distributed in decreasing order from 0 to 6.

claim_amount (Continuous)

Claim amount is the amount claimed by a customer in an event of accident. The maximum amount claimed is around $1.5Million with median value $8900 and mean $48K.

fraud_claims (Categorical)

As per dataset, the number of fraudulent claims are 1 and 2 with the majority of the records being non-fraudulent.

credit_score (Continuous)

The value of credit score ranged from 0 to 784 with median and mean customer credit score of 725 and 670.

annual_premium (Continuous)

The annual premium is the most recent premium paid by the customer. The (min,max) value of premium is between 647.54 to 3199.45. The (median, mean) value of premium is in the range of (771.99,1128.644).

acceleration_x, acceleration_y, acceleration_z,acc_abs_cumm,acc_avg (Continuous):

- The accelerations at x, y, z are given in the acceleration_x,y,z column. We also have an absolute and cumulative absolute value of accelerations.
- We re-engineered the absolute average acceleration values to derive peaks in the jerk/jolt vector.

High Value Customer Definition

We have defined 'High Value Customers' as customers who contribute to 80 percent of the company's positive net revenue.[2] Net revenue is defined as the difference between the total premium paid by the customer over the lifetime of the policy and the claim amount.

To calculate the total positive net revenue, we have considered the subset of customers with positive net revenue.

Assumptions

The annual premium provided in the dataset is the cumulative annual premium of the current year.

Assumption 1 : Hence, we have assumed that the annual premium paid by a customer increases every year @ 5% P.A. over the lifetime of the policy[1]. This has been taken into consideration for calculating the total premium paid by the customer over the lifetime of the policy.

Assumption 2 : We have assumed that the annual_premium paid by the customer is not affected by whether or not a customer has made a claim or not.

Change factor has been calculated by dividing the total premium paid by the customer without considering 5% annual increase to the total premium paid by the customer with 5 % annual increase in premium.

Change_Factor = Total Premium (without 5% annual increase) ÷

Total Premium (with 5% increase P.A.)

Calculations

For example, for policy number **451970 ,** policy start date is 01/25/2013 and the policy effective date is 01/25/2022.So, this customer has been using this policy for 9 years and the latest annual_premium paid by the customer is 758.59.

Assuming that the annual premium has increased @ 5% over 9 years :

Annual Premium paid by the customer in the period 2021-22: 758.59

Annual Premium paid by the customer in the period 2020-21: 758.59 / 1.05 = 722.46

Annual Premium paid by the customer in the period 2019-20: 722.46 / 1.05 = 688.06

Annual Premium paid by the customer in the period 2018-19: 688.06 / 1.05 = 655.29

Annual Premium paid by the customer in the period 2017-18: 655.29 / 1.05 = 624.09

Annual Premium paid by the customer in the period 2016-17: 624.09 / 1.05 = 594.37

Annual Premium paid by the customer in the period 2015-16: 594.37 / 1.05 = 566.07

Annual Premium paid by the customer in the period 2014-15: 566.07 / 1.05 = 539.11

Annual Premium paid by the customer in the period 2013-14: 539.11 / 1.05 = 513.44

Total Premium = 758.59 + 722.46 + 688.06 + 655.29 + 624.09 + 594.37 + 566.07 + 539.11 + 513.44 = 5661.51

Total Premium (when 5% annual increase is not taken into consideration) = 758.59 * 9 = 6827.31

Change_Factor = Total Premium (without 5% annual increase) ÷

Total Premium (with 5% increase P.A.)

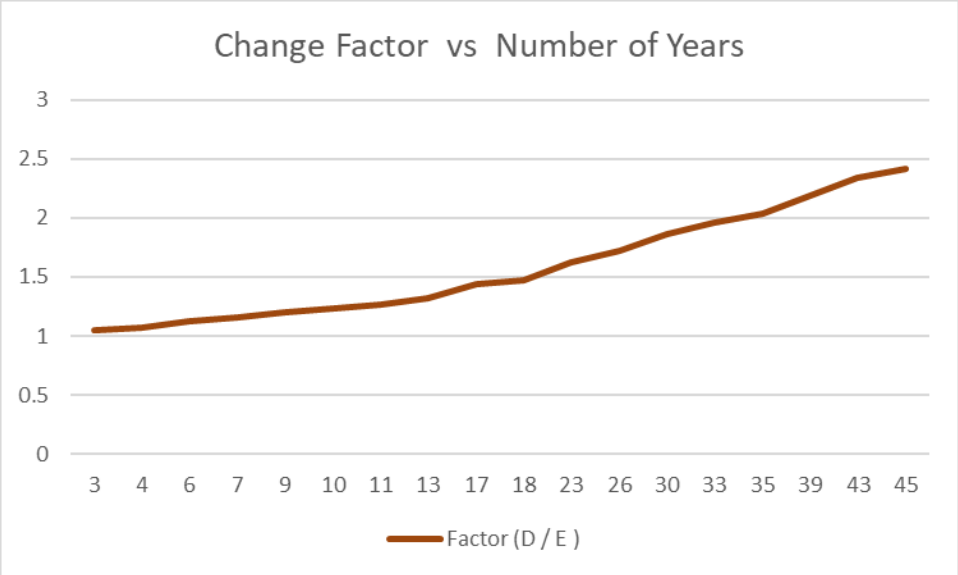
Change_Factor = 6827.31 ÷ 5661.51 = **1.20**

We would use this change factor to arrive at the actual annual_premium paid by the customer over the life time of the policy.

We calculated the change_factors for different policies and plotted change_factors against the number of years (the policy has been active).

Annual_Premium	Years	Total Premium without annual increase	Total with 5% increment over the years	Difference	Change_Factor
1463.88	3	4391.64	4185.83	-205.81	1.049168265
1539.8	4	6159.2	5733.05	-426.15	1.074332162
1545.31	6	9271.86	8235.694	-1036.166	1.125814048
1352.01	7	9464.07	8214.39	-1249.68	1.152133025
758.59	9	6827.31	5661.519	-1165.791	1.205914879
685.55	10	6855.5	5558.31	-1297.19	1.233378491
771.64	11	8488.04	6730.03	-1758.01	1.261218746
718.67	13	9342.71	7088.42	-2254.29	1.318024327
1406.97	17	23918.49	16655.38	-7263.11	1.436081915
1401.08	18	25219.44	17196.94	-8022.5	1.466507414
720.73	23	16576.79	10207.7	-6369.09	1.623949567
1264.49	26	32876.74	19086.14	-13790.6	1.722545261
728.08	30	21842.4	11751.99	-10090.41	1.858612882
731.75	33	24147.75	12295.35	-11852.4	1.963974185
708.07	35	24782.45	12173.77	-12608.68	2.03572517
742.81	39	28969.59	13272.44	-15697.15	2.182687584

750.93	43	32289.99	13834.53	-18455.46	2.334014238
778.54	45	35034.3	14529.71	-20504.59	2.411218118



We used the average change_factor from the plot above to calculate the total_premium paid by a customer.

Total Premium = Total Premium (without 5% annual increase) ÷ Change_Factor

Feature Engineering

We have added more features into the dataset for further classification of customers.

1. **Active_Years**

    - This feature represents the number of years a policy has been active (for each customer)
    - It is obtained by calculating the difference between the policy_start_date and policy_effective_date
    - It indicates for how long a customer has stayed with the same insurance company instead of shopping around

2. **Change_Factor**

    - This feature represents a coefficient that accounts for the change in annual_premium over the lifetime of a policy
    - We have assumed the annual increase in premium to be 5%

Change_Factor = Total Premium (without 5% annual increase) ÷

Total Premium (with 5% increase P.A.)

- We have used Change_Factor to calculate the total premium paid by a customer over the lifetime of the policy

3. **Total_Premium**

    -This feature represents the total premium paid by a customer over the lifetime of the policy

    -It is the product of (Active_Years & Annual_Premium) divided by the Change_Factor

$$Total\ Premium\ =\ (Active\ Years\ *\ Annual\ Premium)/\ Change\ Factor$$

4. **Net_Revenue**

    - This feature represents the revenue generated by the insurance company per policy
    - It is the difference between Total_Premium and Claim Amount

$$Net\ Revenue\ =\ Total\ Premium\ -\ Claim\ Amount$$

5. **High_Value**

    - This is a categorical feature which indicates if a customer is a high_value customer or not
    - A value of '1' represents a 'high_value' customer
    - We have considered customers who contribute to 80% of the positive net_revenue of the company as 'High_Value' customers

Jerk / Jolt / Surge

The dataset provided includes acceleration values at every instant of time in x,y,z direction. Moreover, the dataset also has an absolute sum of x,y,x accelerations and absolute average of x,y,z accelerations for every instant of time.

In physics Jerk/Jolt/Surge is defined as the change in acceleration with respect to time. When there is a huge change in acceleration , it is called **High Jerk**. Intuitively, we can explain this with the example of "spring". If we constantly exert a force on a spring, it stays the same length. But if we let it go , the spring snaps back to its original length.

Usage of acceleration in our analysis:

The rate of change of acceleration can be a reliable predictor to visualize the nature of driver. If there are sudden changes in slopes in his jerk vector, he can be considered as a rash driver.

Computing surge vector:

Step: 1 Rate of change of acceleration is coined as surge/Jolt. Below is the formula of computing Surge vector (S) based on A(acceleration) at time instant (t).

$$S = \frac{A(t2) - A(t1)}{t2 - t1}$$

Step: 2 Compute peak values. The number peaks can be attributed to how rash the driver is.

6. **Positive_Peaks**

- This feature represents the number of times there has been a sudden increase in acceleration , for a driver over the time period under consideration.
- We have used the difference between every consecutive pairs of average acceleration values to calculate the increase / decrease in acceleration. (Jerk)
- The jerk values were then plotted on a graph against time and the number of peaks on the positive half were calculated for each customer to get the number of positive peaks.

7. **Negative_Peaks**

- This feature represents the number of times there has been a sudden decrease in acceleration , for a driver over the time period under consideration.
- We have used the difference between every consecutive pairs of average acceleration values to calculate the increase / decrease in acceleration. (Jerk)
- The jerk values were then plotted on a graph against time and the number of peaks on the negative half were calculated for each customer to get the number of negative peaks.

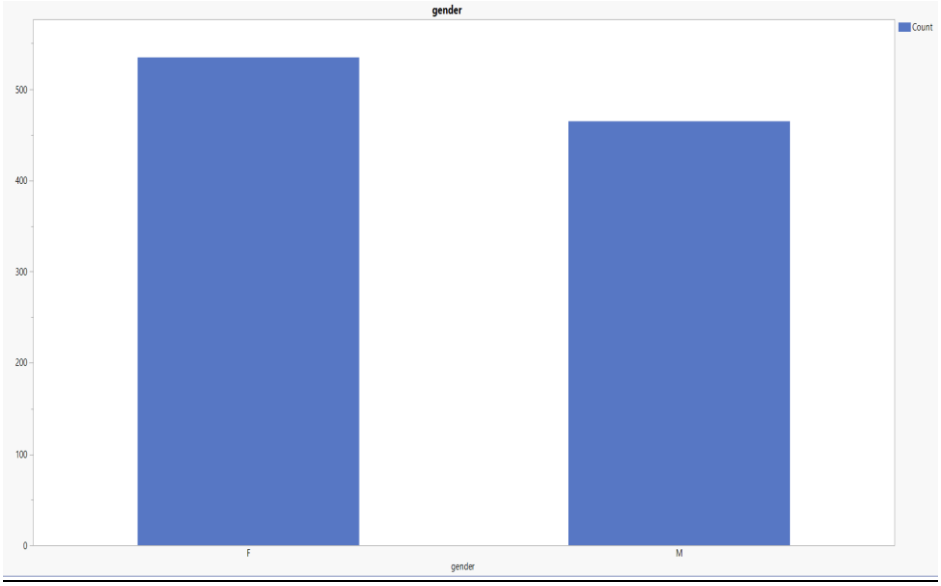
8. **Car_Type**

- We have used a combination of auto_make and auto_model to identify the type of car for each customer
- The cars have been classified into 5 types : Compact SUV , Sedan , SUV , Sports Sedan & Truck
- We have used the below source to identify the type of the car [3]: https://www.cars.com/shopping/?aff=acqgeosem20&KNC=acqgeosem20&utm_source=google&utm_medium=cpc&utm_term=buy%20used%20cars&t_id=kwd-96368416&gclid=Cj0KCQjwl7qSBhD-ARIsACvV1X3uvUfrLJ9e3n5llTTtURLy2eGGnoevBtplSXkkqReooiJ_mZupSToaAk3hEALw_wcB&gclsrc=aw.ds

Analysis

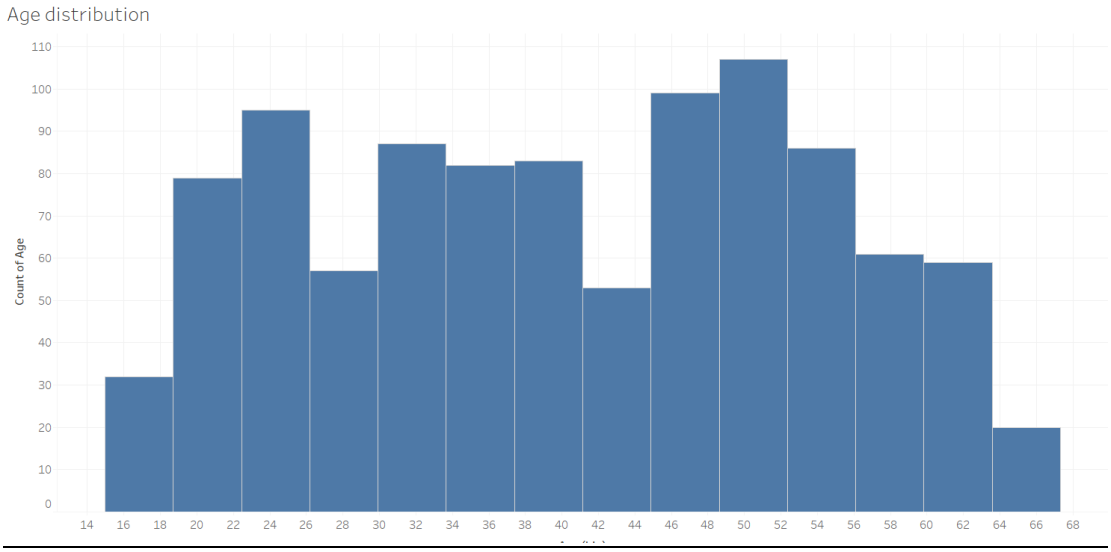
General observations about the dataset

1) Distribution of Gender



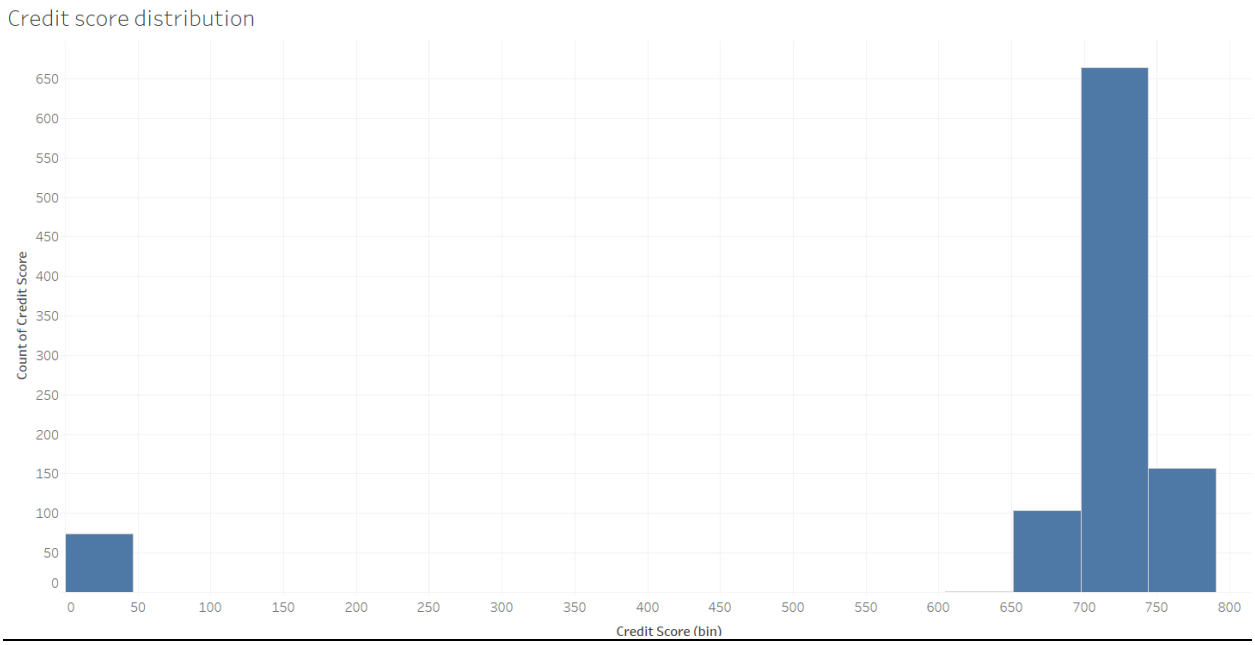
- There are 535 Females and 465 Males in the dataset provided

2) Distribution of Age



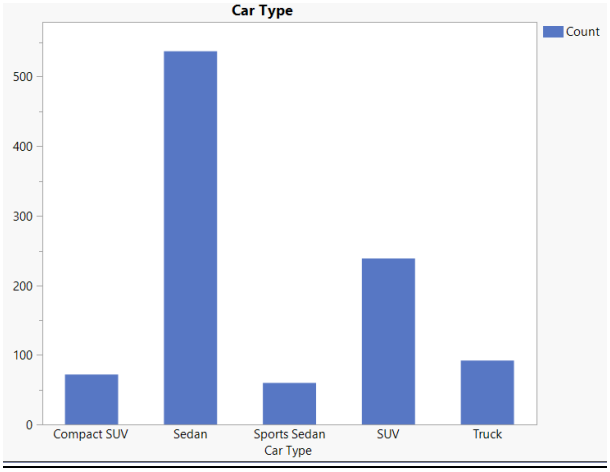
- The age of the customers is distributed between 18 years & 64 years
- The median age of the customers is 40 years
- More than 50% of the customers are over 40 years of age

3) **Distribution of Credit_Score**



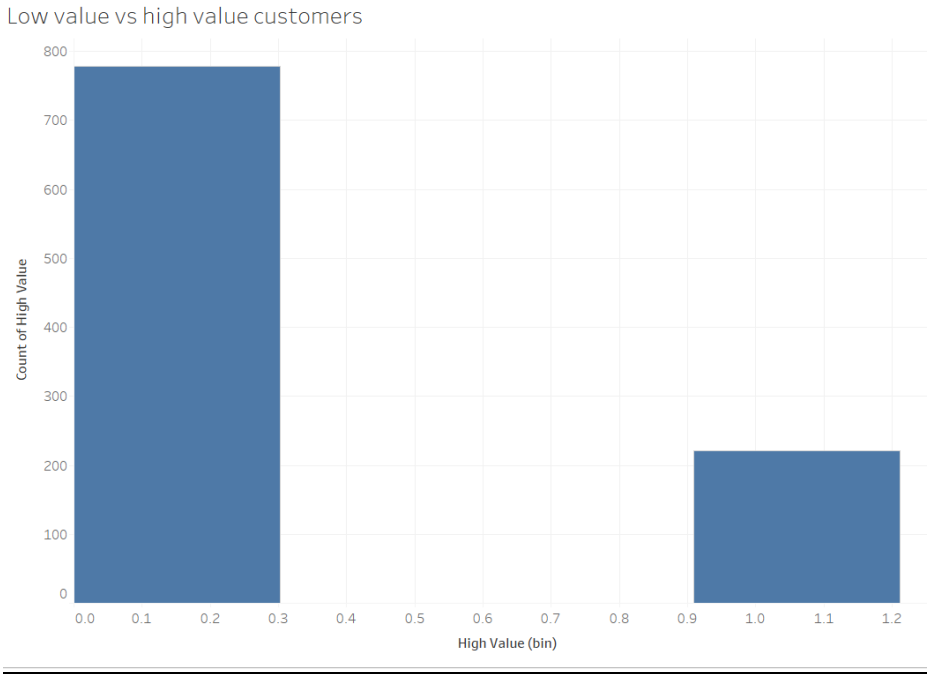
- The Credit_Score of the customers is distributed between 0 and 783
- The median credit score of the customers is 725
- More than 50% of the customers have credit scores above 725

4) **Distribution of car_type**



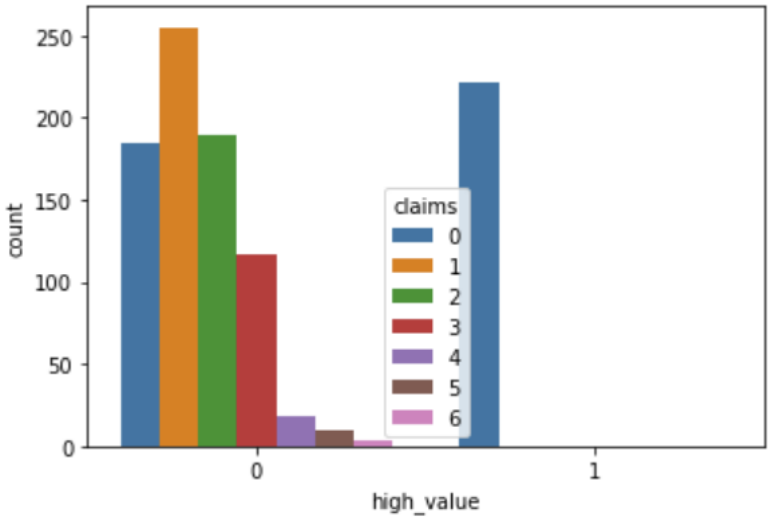
- We can see that the majority of customers own Sedans (more than 500 customers)
- The second most popular car_type is SUV with about 250 customers owning an SUV

Classification of High Value Customers



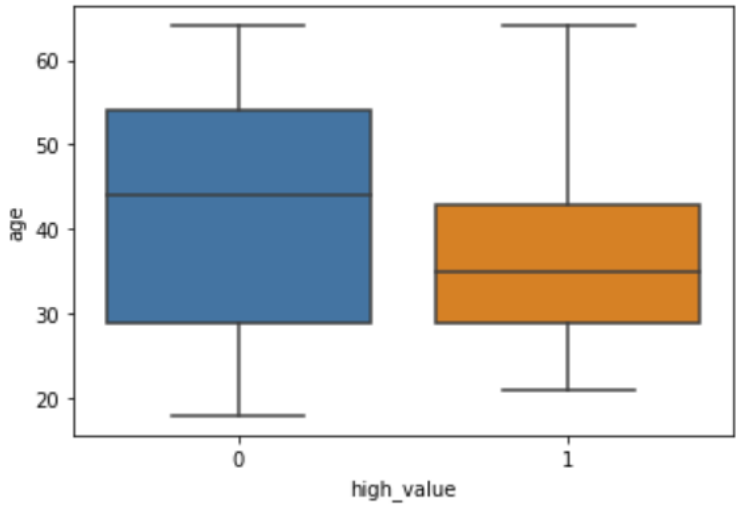
- Based on our analysis ,we found that there are 221 High_Value customers in our dataset based on their contribution towards the total positive revenue of the insurance company

Number_of_Claims Vs High_Value Customers



- The subset of 'High_Value' customers have made 0 claims throughout the lifetime of the policy.
- The rest of the customers not classified as 'high_value' have made claims ranging from 1 to 6. A majority of these customers have 1 or 2 claims throughout the lifetime of the policy while only a very small percentage of them have made 4 to 6 claims.

Age Vs High_Value Customers

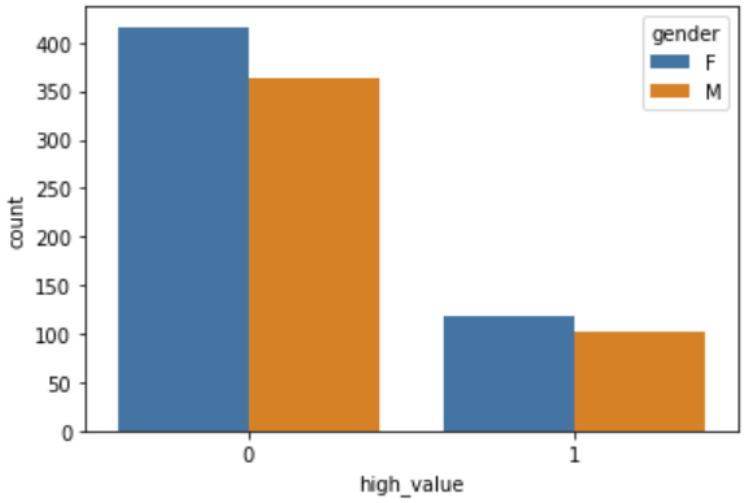


- We can see that the boxplots for the two categories are slightly staggered
- The median age for 'high_value' customers is lower than the other category
- Median age for 'high_value' customers is 35
- Median age for customers not classified as 'high_value' is 44.

- Summary statistics are as shown below :

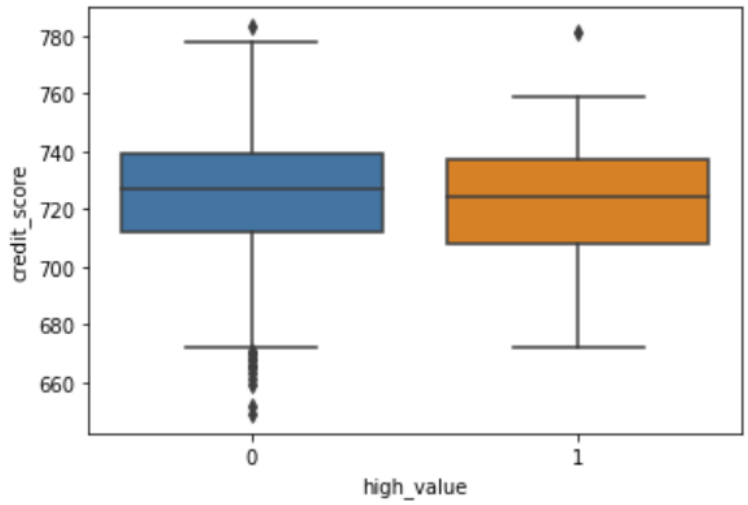
high_value	count	mean	std	min	25%	50%	75%	max
0	779.0	41.652118	13.999299	18.0	29.0	44.0	54.0	64.0
1	221.0	36.389140	9.576613	21.0	29.0	35.0	43.0	64.0

Gender Vs High_Value Customers



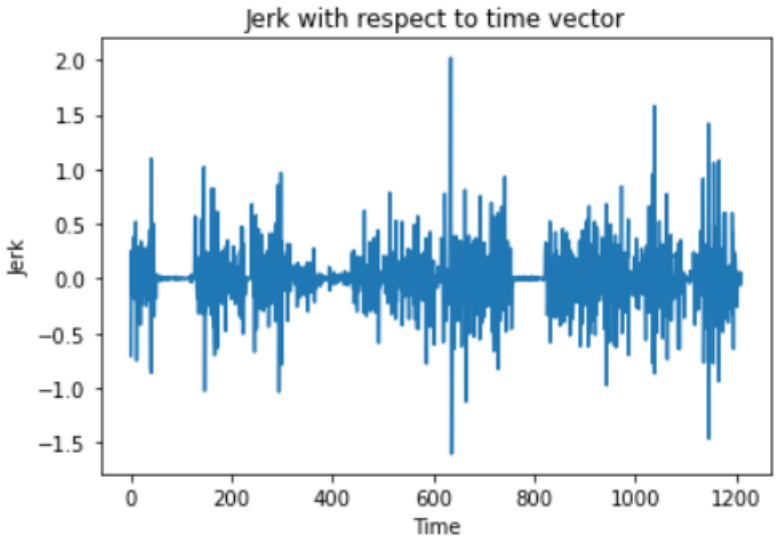
- The ratio of number of females to number of males is slightly higher for 'high_valued' customers when compared to the category of customers that are not classified as 'high_valued'

Credit_Score Vs High_Value Customers



- We see that there is no major distinction between ‘high_value’ customers and customers not classified as ‘high_valued’ in terms of credit_score
- The boxplots for both the categories overlap each other
- Since credit_score depends on the creditworthiness of the customer ,it does not contribute towards deciding whether a customer is ‘high_value’ or not

Jerk with Respect to Time

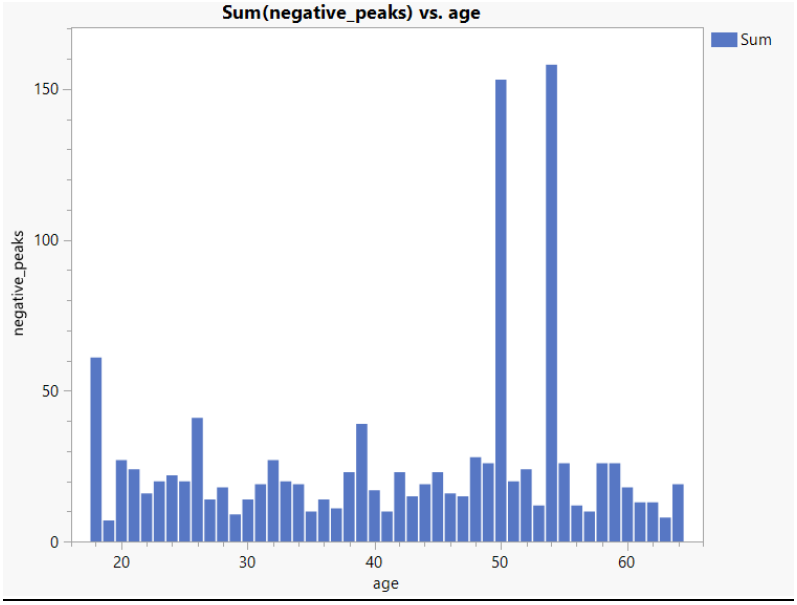


- We have considered Jerks with magnitudes more than +2 or less than -2 to be peaks

Conclusions

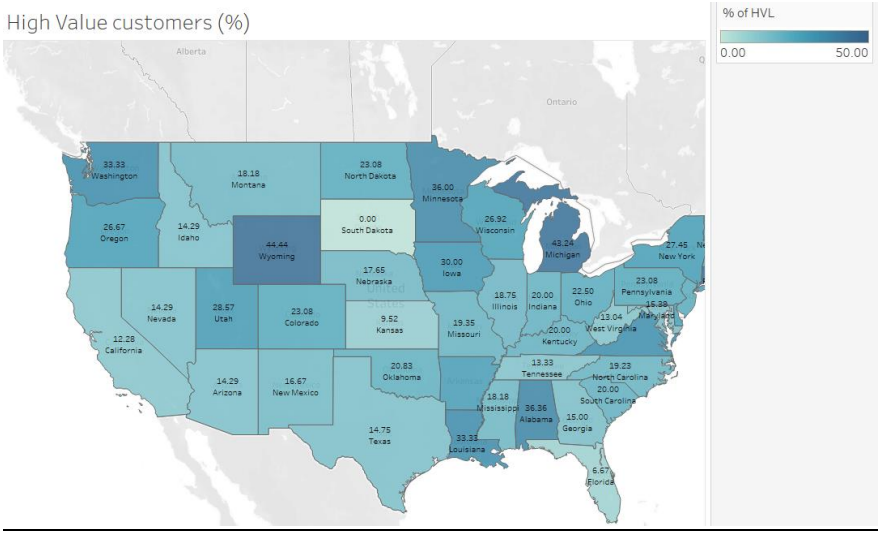
Conclusion 1

- You can see from the below chart that the number of sudden jerks are higher among the ages 50 to 55 and for ages below 20
- There is less number of jerks for people below 50 years because the response rate is faster for lower aged population which results in early braking avoiding sudden decreases in acceleration (sudden jerks).
- Since sudden jerks (sudden decelerations) can be attributed to more number of accidents , which could lead to more number of claims, the higher aged population generates lower revenues for the company.
- Also since the customers below 20 years of age are new to driving , there are higher chances of accidents among this age group as well leading to more number of claims and hence lesser revenue
- The mid 50% of the 'high_value' customers lie in the range of 29 to 43 years with 0 accidents , corresponding to 0 claims
- They have very less number of spikes in acceleration as is evident from the chart below



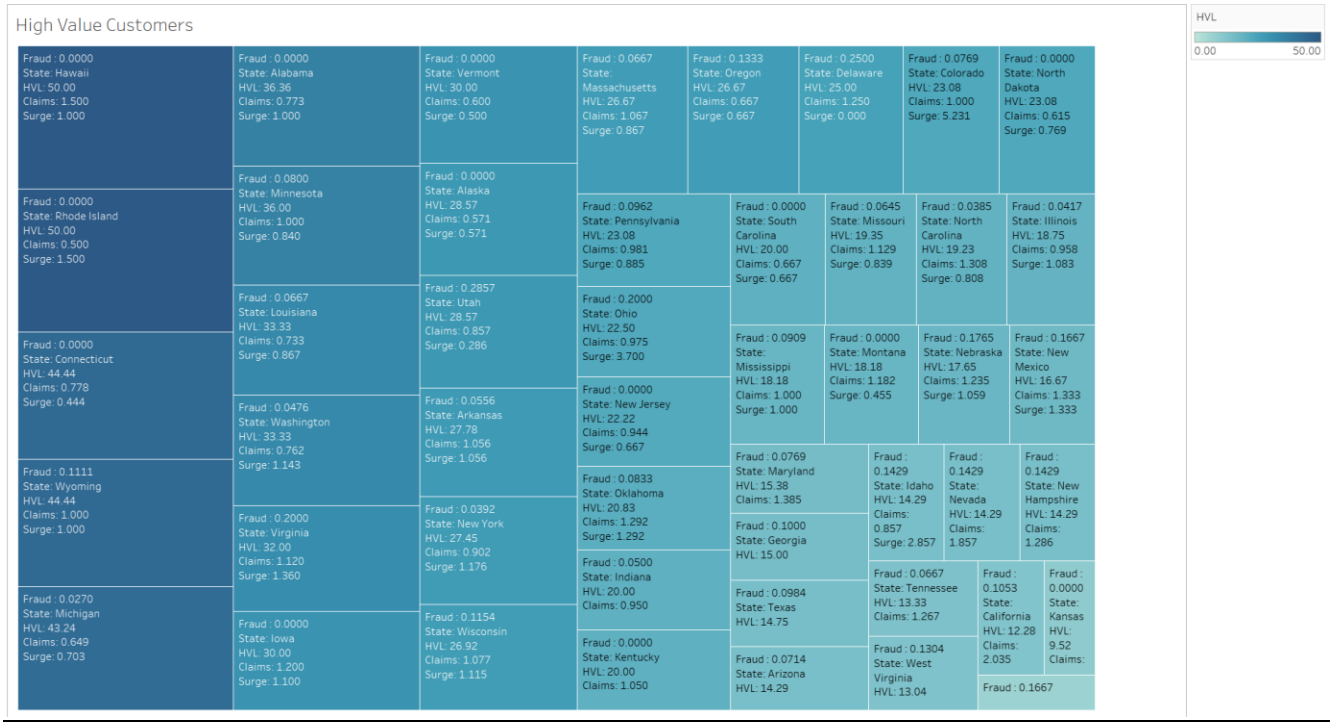
Conclusion 2

Distribution of High Value Customers (HVL)



- We can conclude that the percentage of ‘high_valued’ customers is the highest for the states Wyoming , Michigan , Alabama and Minnesota.
- Although the number of customers from these states are much less when compared to states like New York , California and Texas , these customers contribute the most towards the company’s revenue

Conclusion 3



- In states Hawaii, Rhode Island, Connecticut, Wyoming, Michigan with high concentration of high valued customers, the Surge (Jerk), claim, fraud factors are minimal. This means that high valued customers are careful at driving, which leads to less to no claims.
- In low performing states like Washington, Arkansas, Iowa, Colorado the surge factor is more than one leading to more claims in most of the states.

Recommendations

1. **Customer Retention Strategies**

- Since more than 50% of the high_value customers fall between the ages of 29 & 43 and most people start owning houses at this age , additional home insurance could be offered in combination with the car insurance at attractive prices / premiums . This could be a huge incentive to the 'high_value' customers and can encourage them to stay with Blue Buffalo Insurance
- Since 'high_value' customers have made no claims during the lifetime of their policy , they should be offered a no-claim bonus. This bonus could be offered at the time of every renewal. The bonus amount could be calculated based on the number of years the customer has been associated with Blue Buffalo Insurance.
- Direct maximum resources towards the high valued customers by providing regular check-ins and personalized support to ensure customer satisfaction

2. **Customer Acquisition Strategies**

- Michigan, Wyoming, Alabama, and Minnesota account to a mere 9% of the entire customer base but consist of a huge percentage of the high valued customers. Targeted marketing should be done to acquire more customers from these states.

- Utilize the list of high-valued customers in order to get referrals of other customers which leads to – low acquisition rate and potentially an increase in the number of high valued customers.

References

1. https://www.mckinsey.com/industries/financial-services/our-insights/global-insurance-pools-statistics-and-trends-an-overview-of-life-p-and-c-and-health-insurance
2. https://tallyfy.com/high-value-customer/#:~:text=Those%20who%20contribute%2080%25%20of,value%20customers%20are%20important%20too
3. https://www.cars.com/shopping/?aff=acqgeosem20&KNC=acqgeosem20&utm_source=google&utm_medium=cpc&utm_term=buy%20used%20cars&t_id=kwd-96368416&gclid=Cj0KCQjwl7qSBhD-ARIsACvV1X3uvUfrLJ9e3n5llTTtURLy2eGGnoevBtplSXkkqReooiJ_mZupSToaAk3hEALw_wcB&gclsrc=aw.ds