ML: CIA - 2

Churn Analysis On Life Insaurance.

- Name: Hardik Shah
- Register No.: 2241131
- Class: 6 BCA - B

• Dataset description:

This dataset provides a comprehensive overview of customer churn in the life insurance sector, containing 1000 records with 11 features.

- Features:

1.) customer name.
2.) Customer address.
3.) Company Name.
4.) Claim Reason.
5.) Data confidentiality (The level of confidentiality of the data)
6.) Claim Amount.
7.) Category Premium.
8.) Premium Ratio (ratio of claim Amount to the premium amount)
9.) Claim Approved.
10) BMI.
11) Churn (Target variabel).

• The Target variable "Churn", identifies whether a customer has left the Insurance policy or not.

•

customer name	customer address	company name	claim Reason	Data confi-dentiality	claim Amount	category Premium	premium ratio	claim Approved	BMI	churn
Tony	xxx...	WHtP	Travel	Low	377	479	0.075	NO	21	yes
Nicole	xxx...	MG	Medical	Medium	14089	14390	0.100	yes	24	yes
Linda	xxx...	HP	Phone	High	1875	1875	0.124	yes	21	NO

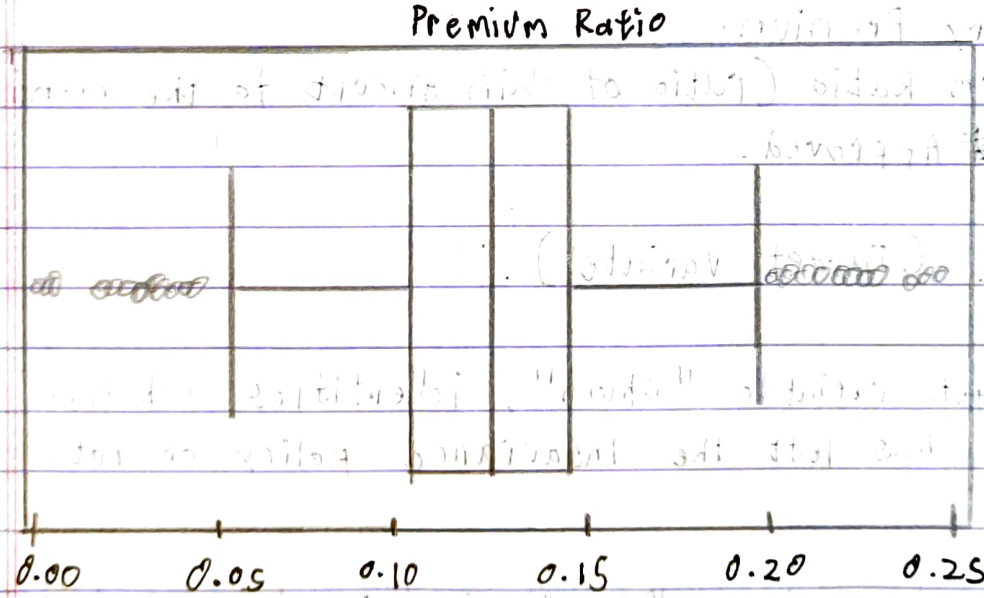
* colomns like; 'customer Name', 'customer Address', 'company name' & 'BMI' have been dropped as they are irrelevant.

* The column 'BMI' has been dropped as it is highly correlated to the Target variable, which is throughing off the analysis.

• EDA performed:

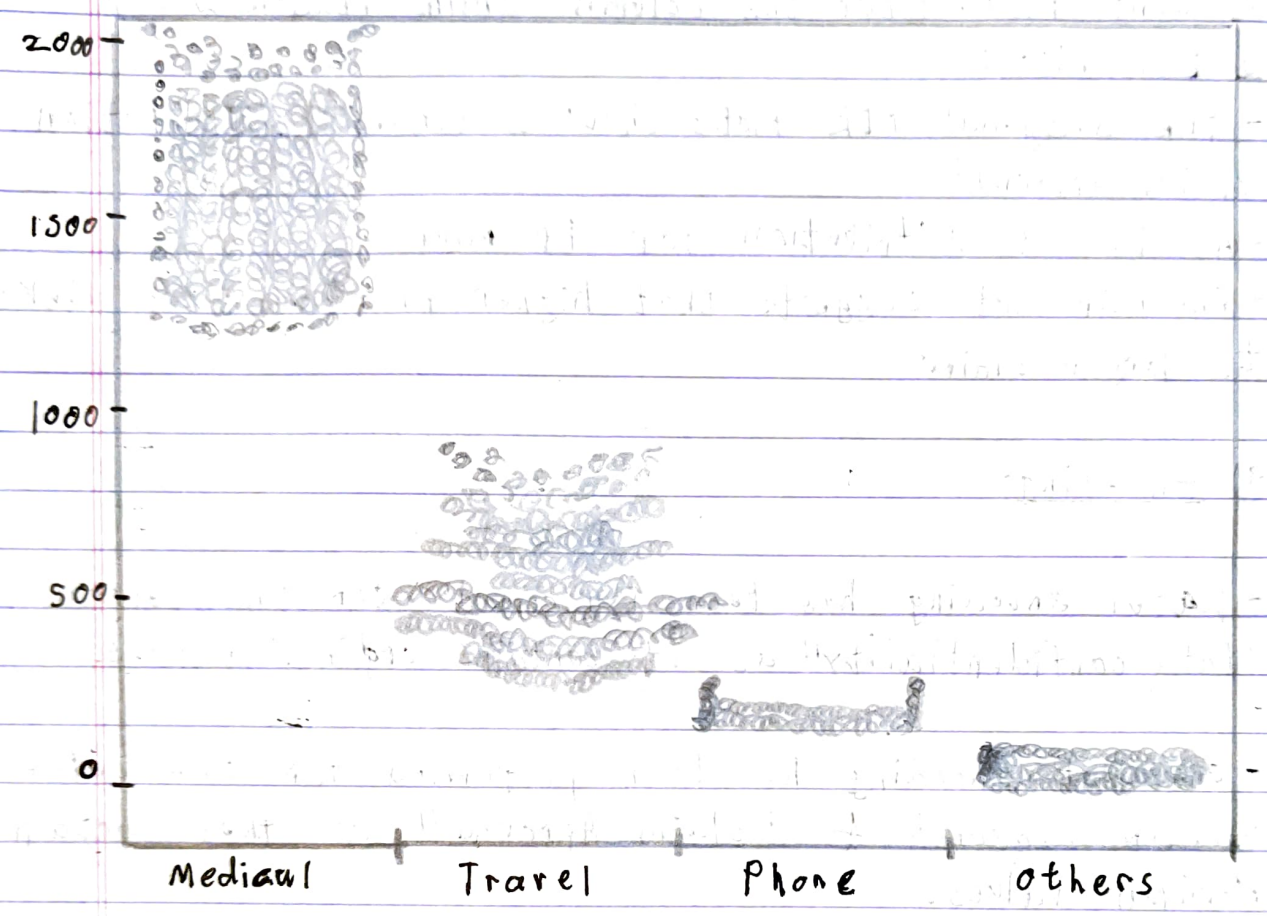
1.) Rows with Null values have been discarded.
2.) Dataset was checked for duplicate rows.
3.) 'churn' was renamed to 'Target' for easier Analysis.
4.) Outlier Detection ('Premium Ratio' was the only column with outliers).

- Box plot:



- There are multiple outliers in the column 'Premium Ratio'. Hence we need to get rid of them.
- The IQR Method was used to remove outliers.

5) Swarm plot: (For the column 'claim Reason').



- Interpretation:

1.) Medical claims have the highest variability & the largest claim amount. This might require a higher risk nanagement.

2.) Travel claims exhibit a mid-range cost pattern but are more varied than Phone & other claims. This is more moderate.

3.) Phone & other claims are associated with consistently low claim amount. They represent minimal risk with low-cost claims.

6) Voilin Plot. (For the column claim reasons).
7) Count Plot. (For the column claim reasons).
8) Pair Plot:
- The diagonal KDE plots show a binomial distribution for 'claim amount'
- A skewed distribution for Premiums.
- The pair plot suggests that higher premium are linked to larger claims

9) Encoding:

- Label encoding has been performed for the column "Data confidentiality" as it contains ordinal values.

- One-Hot encoding has been performed for the columns "Claim Reasons" & "Claim Aproved" as they contain Nominal values.

- Binary Encoding has been performed on the column 'Targe' (Churn).

10) Standard scaling: (all values were scaled using the StandardScaler() method (-1 to 1)).

11) Heat map:

Data confidentaily	1	0.83	0.87	0.027	-0.0052
claim Amount	0.83	1	0.96	0.24	0.0033
Category Premium	0.87	0.96	1	0.021	-0.0039
Premium Ratio	0.027	0.24	0.021	1	-0.02
Target	-0.0052	0.0033	-0.0039	-0.02	1

Data confidentiality claim Amount Category Premium Premium Ratio Target

• Interpretation:

- category premium & claim Amount have a very high corelation (0.96). Indicates that higher premiums are associated with higher clain amount
- Data confidentiality is highly corelated with category Premium & clain amount, soggesting that this data might be related to high profile clients.
- The strong corelation between claim amount & category premium soggests potential multicollinearity.

• ML Analysis:

• 1) KNN: K nearst neighbour

- Knn is a supervised learning algorithm, used for classification & regression that classifies a data point based on the majority class of its K nearst neighbors.

- Key characteristics:

i) Lazy Learning: No traing phase.
ii) Instance Based: classifies based on similarity.
iii) Distance Based: Euclidean Distance.

- Algorithm

i) Choose the number of neighbors 'K'.
ii) Calculate distance between the query point & all data points.
iii) sort the distance & select the 'K' nearst neighbors.
iv) Perform majority voting (classification) or take the average (Regression).
v) Return the predicted lable / value.

- Chrun Analysis:

- Using the Error Rate vs. K value plot we select the best 'K' value (minimum Error Rate).
- According to the Error Rate plot the best value for 'K' is 36.

- Results:
- confusion matrix:

	0	1
0	2	58
1	0	115

∴ Accuracy = 66.85 %.

- Interpretation:

i) class 0 (Negetive class)

- Precision = 1.00: The model predicted class 0 perfectly, but it does not account for the fact that it rarely predicted class 0.

- Recall = 0.03: only 3% of actual class 0 instances were correctly predicted as class 0. This indicates poor sensitivity for this class.

- F1 score = 0.06: A low F1 score reflects poor performance of class 0.

- Class 1 (positive class)

- Precision = 0.66 : when the ~~class~~ model predicts class 1, 66% of those predictions were correct.

- Recall = 1.00 : The model identified all class 1 instances correctly.

- 2) Decision Tree:

i) Using Gini Index:

- Using the Gini criteria the model selected the column 'Claim Amount' as the Root decision node.

- Results:

- Accuracy: 66%
- Precision: 0.64
- Recall : 0.66
- F1 score : 0.55

- Confusion Matrix:

	0	1
0	3	57
1	2	113

ii) Using Entropy criteria:

- The Entropy model also selected 'clain amount' as the root decision node.

- Results:

- Accuracy: 66%
- Precision: 0.64
- Recall : 0.66
- F1 score: 0.55

• Interpretation:

- The decision tree classifiers (Gini & Entropy) achived moderate accuracy (66.29%) but showed significant bias towards class 1
- Both models had poor performance for class 0 with low recall & F1 score, while class 1 achived a high recall (98%) & a stronger F1 score.
- The AUC scores (Gini = 0.54 & Entropy = 0.51) indicates week discriminatory power.
- The dominant feature in both classes is claim Amount, followed by premium Ratio.

• Comparative Analysis: (conclusion).

- Overall Accuracy: Both models achieve similar accuracy, indicating that they are moderately effective overall.

- Class level Analysis: KNN achieves perfect precision for class 0, but its recall is extremely low. Decision tree has a slightly better recall but low precision. Both models perform well for class 1, with high recall + F1 score.

- Both models have weak discriminatory power, with score close to 0.5, indicating that neither is reliably at separating classes.

- To improve the performance, particularly for class 0, it is essential to address the class imbalance through data rebalancing techniques.