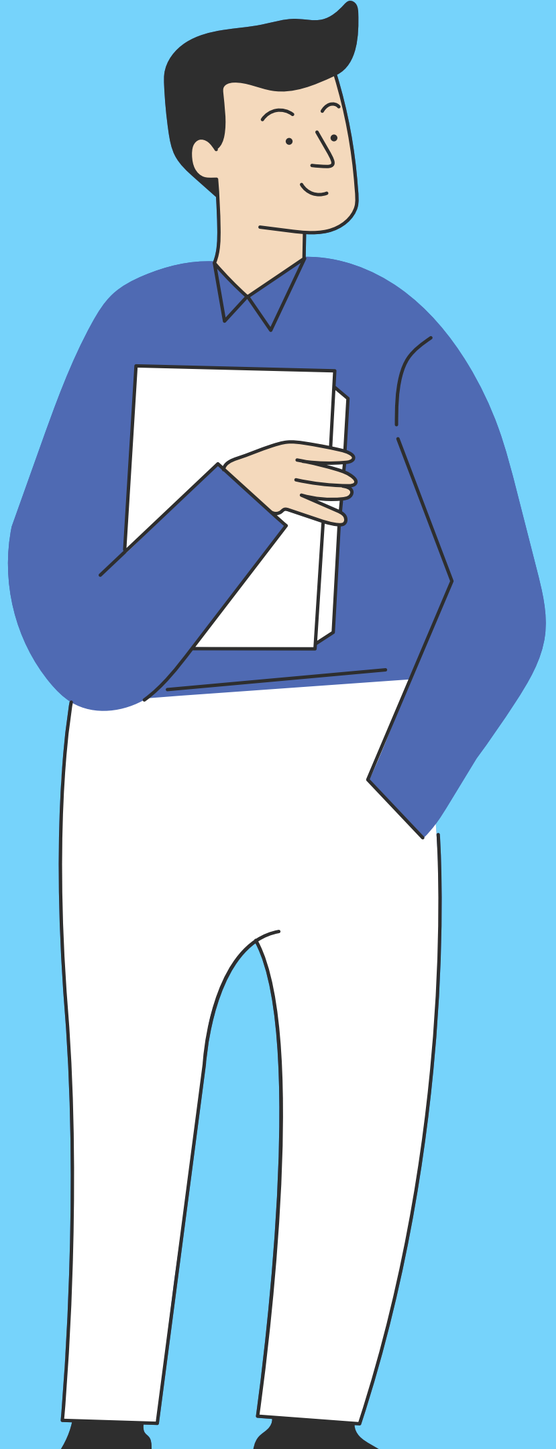


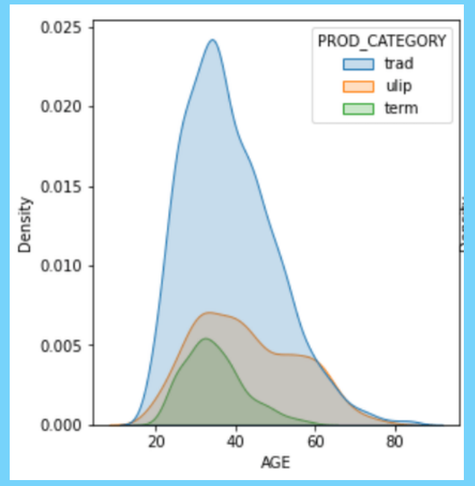
SOLUTION PRESENTATION OF BAJAJ ALLIANZ HACKATHON

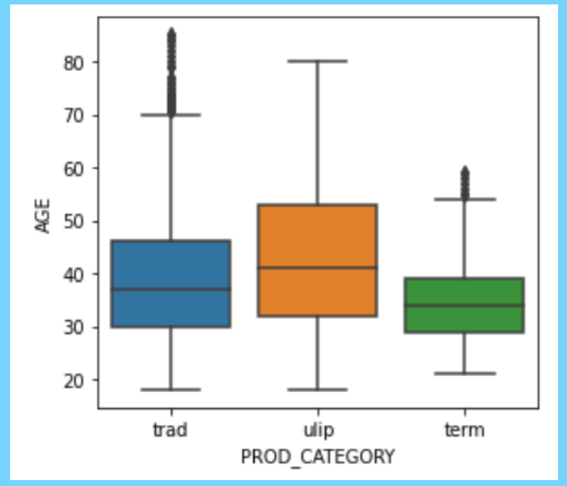
By
Bibekananda sahoo

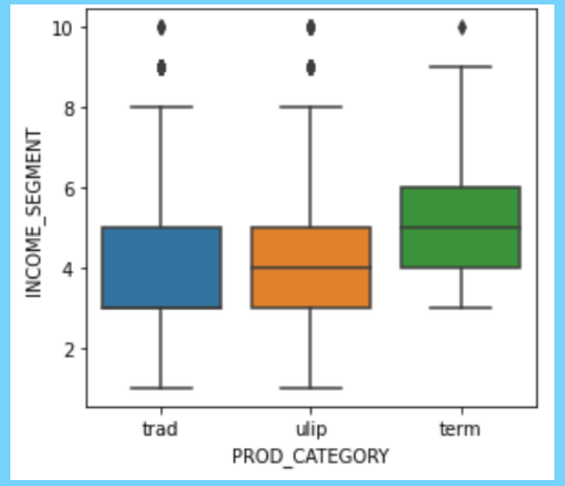
Data Understanding & EDA

Overview

The Training data has a total of 4500 observations with 10 features out of which 5 are categorical. The Test data has 500 observations. The target feature is PROD_CATEGORY which has information about the types of insurance policy.

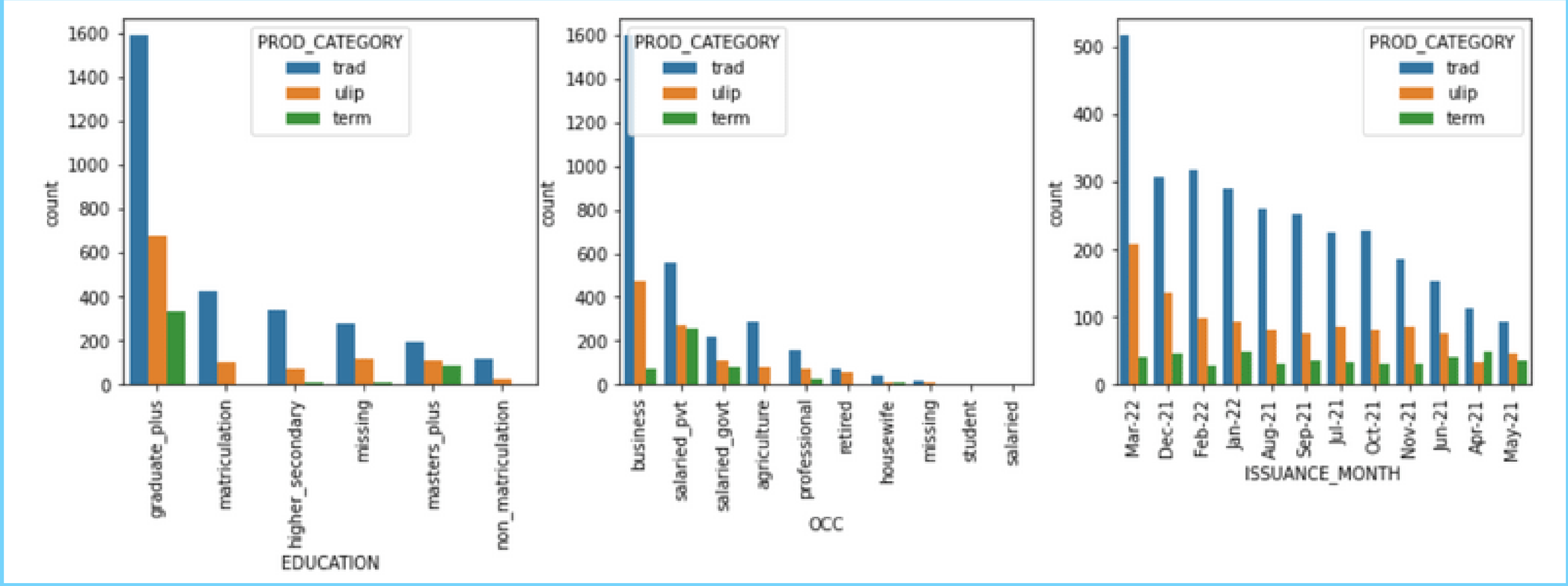






- Clients taking Trad and Ulip product categories have an age range from 20 to 80 and in Term, the category age ranges from 20 to 60 also Age has a slightly positively skewed distribution.
- The mean Income segment for the product category Term varies from 3 to 10 and has a mean of 5 and for the Trad and Ulip categories, the mean income segment is 5 and 4 respectively.

Data Understanding & EDA



- Peoples are more like to buy insurance in the months of march, Dec, Feb, Jan. and in the month of April clients mostly prefer the product category Trad and Term.
- Most of the clients are comes from occupations having business and salaried privet and education having graduate plus.

Data Preparation

Dealing with missing data:

- There are some missing values present in the Age and Pincode column for the Age column I replace missing values with mean and for Pincode I dropped those rows because the percentage of missing values for Pincode is very low.
- Also, there are some -99 values present in column INCOME_SEGMENT so I replace them with mean.

Dealing with Outliers: Mostly the outliers are in continuous nature there are no such outliers present so detached from the body of the boxplot so I leave them as it is.

Scaling and Encoding: For scaling the numerical columns I used a Standard scaler and used one hot encoding for categorical columns.

Feature Engineering

New Features Created -

Stock Market Index

- Extracting data for all indices in NSE from April 2021 to March 2022 Along with Gold price data (INR per troy ounce).
- Where 1 Troy ounce = 31.1035 gram

Unemployment Rate

- Extracting the Estimated Unemployment Rate and Estimated Labour Participation Rate from April 2021 to March 2022 from

State pincode mapping

- Fetching the state name from the respective Pincode to make a new feature called State.

Model Building & Evaluation

Hyper-Parameter Tuning and choosing the best model-

	Model	Best_Score	Best_Parameter
8	GradientBoostingClassifier	0.680527	{'n_estimators': 100}
0	Logistic_Regression	0.663856	{}
5	RandomForestClassifier	0.654735	{'criterion': 'entropy', 'n_estimators': 100}
2	XGBClassifier	0.654068	{}
1	SVC	0.651626	{'C': 1.0, 'gamma': 'scale', 'kernel': 'linear'}
4	KNeighborsClassifier	0.635395	{'n_neighbors': 6}
6	BernoulliNB	0.631621	{}
3	Decision_Tree	0.591816	{'criterion': 'gini'}
7	GaussianNB	0.195424	{}

- GradientBoosting and Logistic Regression are seems best model for this data

Feature Selection-

- using RFE(Recursive Feature Elimination) to get the top features.
- Some of the best features are:-
- 1. 'AGE'
- 2. 'INCOME_SEGMENT'
- 3. 'Gold_price'
- 4. 'EDUCATION'
- 5. 'OCC'
- 6. 'ISSUANCE_MONTH'
- 7. 'State'

Model Building & Evaluation

Logistic regression model

- Overall accuracy comes in the logistic regression model was 0.68
- For predicting Term the f1-score was 0.32 and for Trad and Ulip the accuracy was 0.80 and 0.25 respectively.

	precision	recall	f1-score	support
0	0.50	0.23	0.32	450
1	0.70	0.94	0.80	2945
2	0.50	0.17	0.25	1103
accuracy			0.68	4498
macro avg	0.57	0.44	0.46	4498
weighted avg	0.63	0.68	0.62	4498

Gradient boosting model

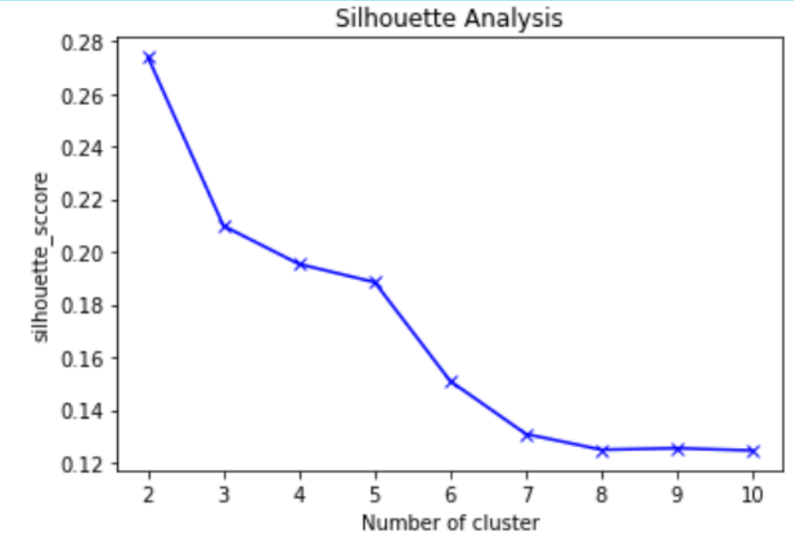
- The overall accuracy comes in the Gradient Boosting model was 0.72
- For predicting Term the f1-score was 0.49 and for Trad and Ulip the accuracy was 0.83 and 0.36 respectively.

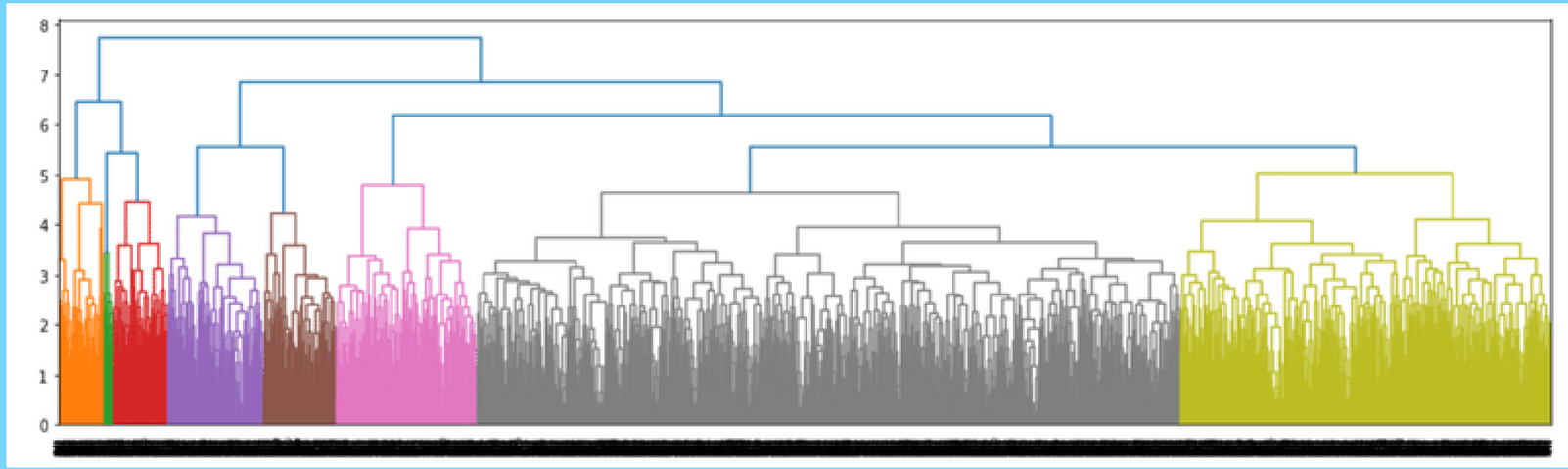
	precision	recall	f1-score	support
0	0.60	0.41	0.49	450
1	0.73	0.95	0.83	2945
2	0.73	0.24	0.36	1103
accuracy			0.72	4498
macro avg	0.69	0.53	0.56	4498
weighted avg	0.72	0.72	0.68	4498

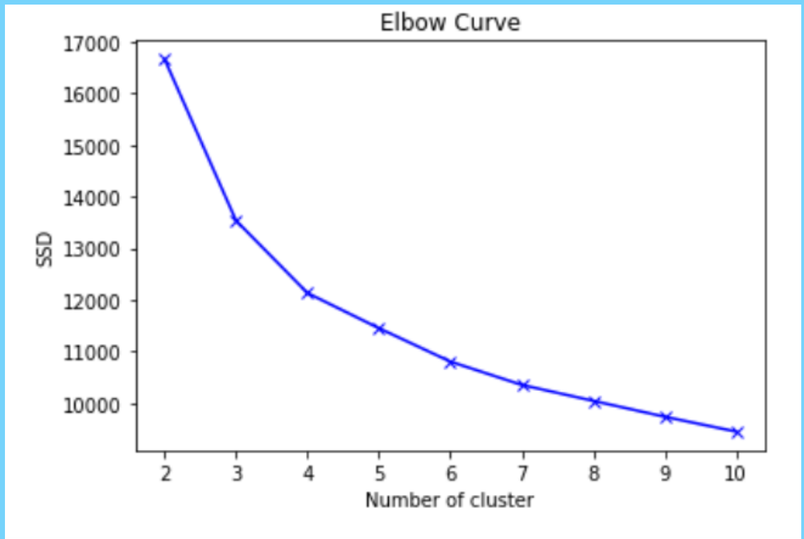
Overall Gradient boosting model is performing better so I take gradient boosting as final model for prediction.

Clustering

Finding Optimal Cluster



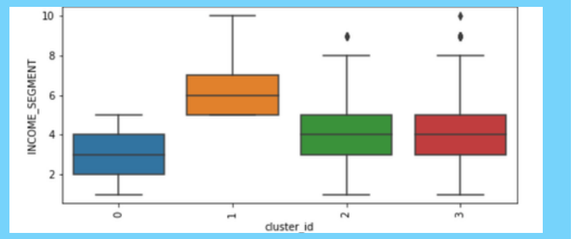


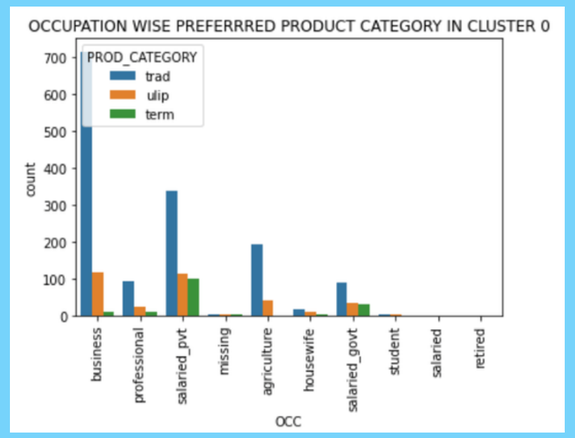


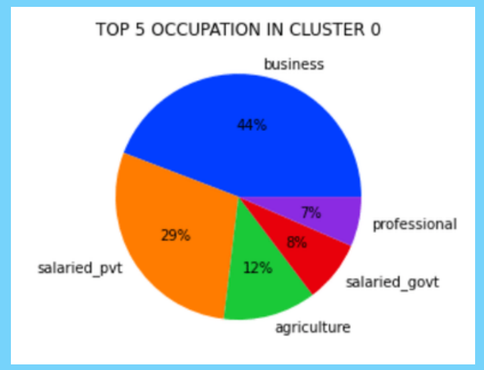
- After analyzing the Silhouette score, Elbow curve, and hierarchical dendrogram four clusters are seems good for this model.

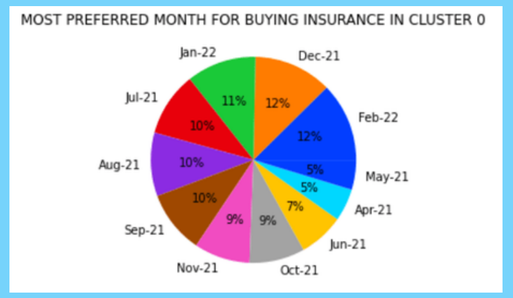
Cluster Summary

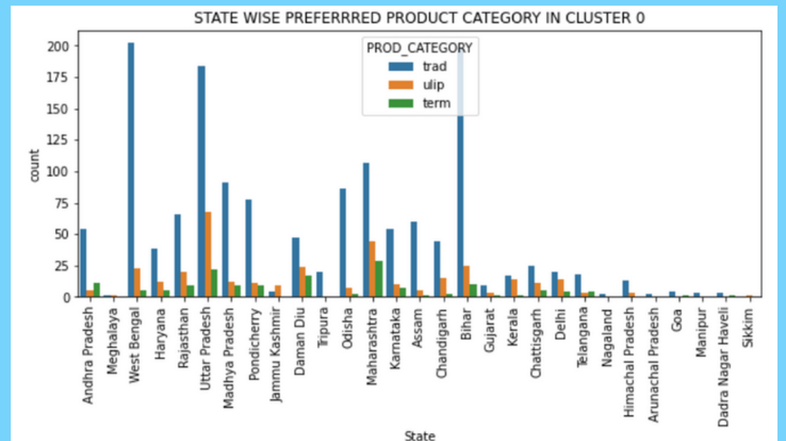
Cluster 1







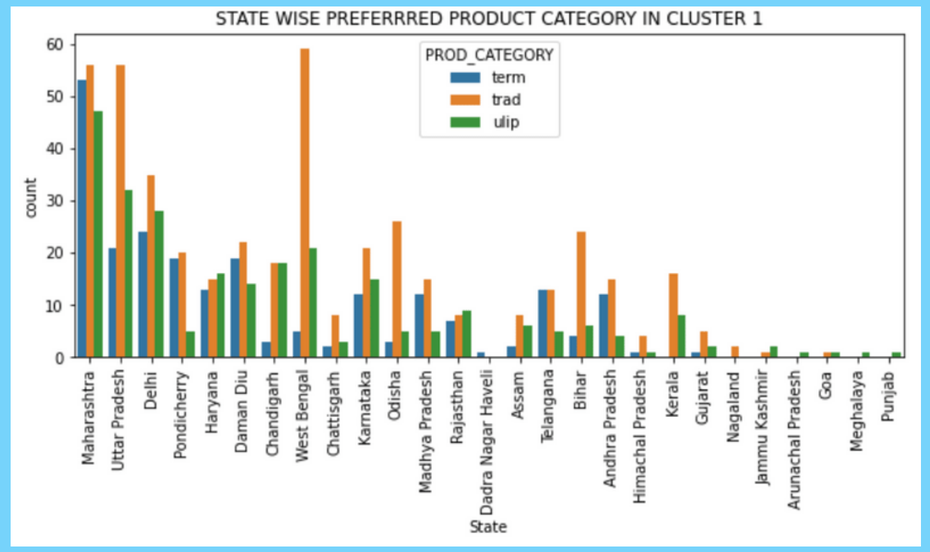


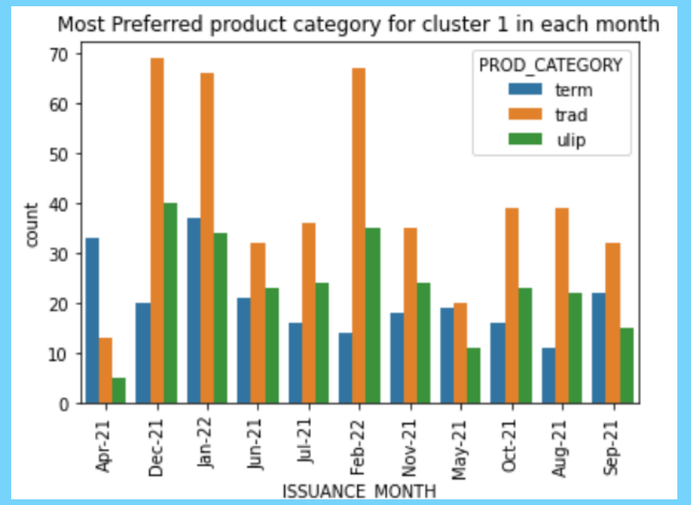


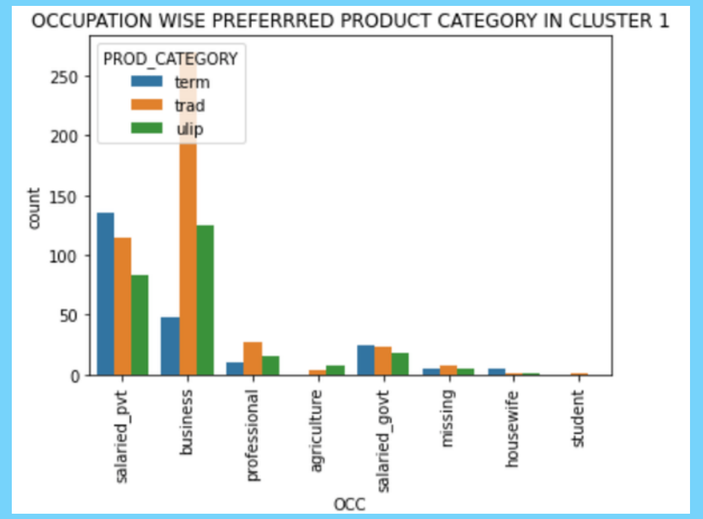
- Clients from this Cluster are given first preferences to Trad insurance and then Ulip.
- But from states like Andhra Pradesh and Telangana clients are giving second preference to the Term product category.
- For this cluster clients mostly comes from northeast states like Utter Pradesh, Bihar, and West Bengal.
- Clients are mostly buying insurance in the months of February, January, and December
- For this cluster, the ages vary from 20 to 50 and the mean age was 33.
- Clients from these clusters mostly come from the low-income segment category.
- Most clients are having Education type graduate plus and matriculation.
- For this cluster, the clients are mostly having occupations such as Business, Salaried private, and Agriculture.

Cluster Summary

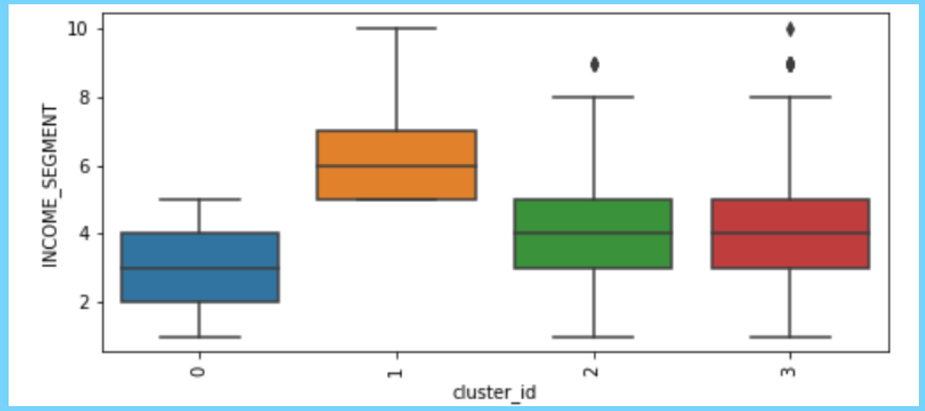
Cluster 2





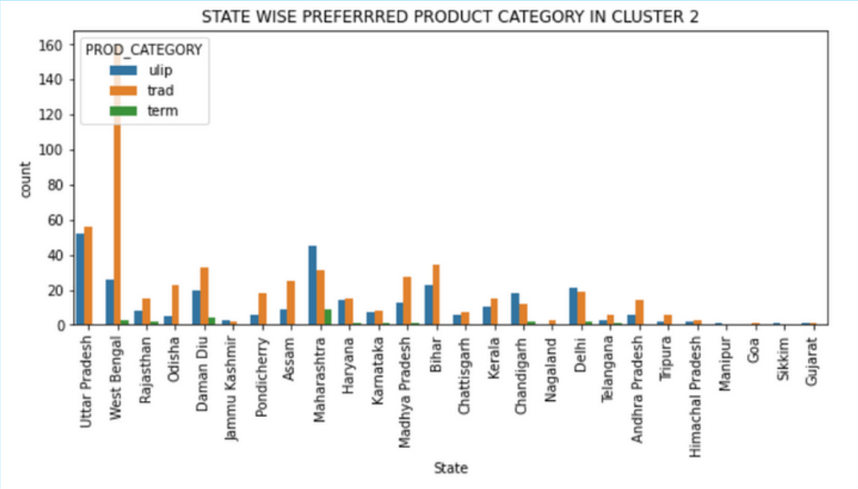


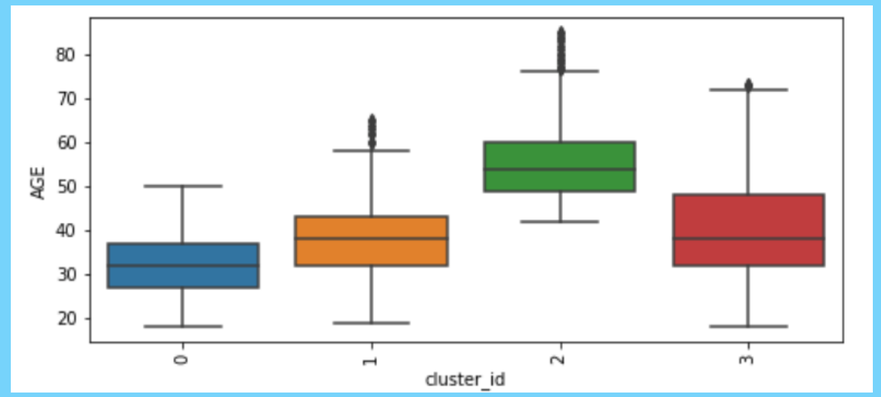
- For this cluster, the ages vary from 20 to 60 and the mean age was around 40.
- Clients from these clusters mostly come from the high-income segment category.
- Clients from this Cluster are given first preferences to Trad insurance and then mostly in April month clients preferred to buy term insurance.
- In states like Hariyana, Chattisgarh, and Rajasthan clients give first preference to the Ulip product category.
- For this cluster clients having occupation salaried privet give first preference to the Term product category.

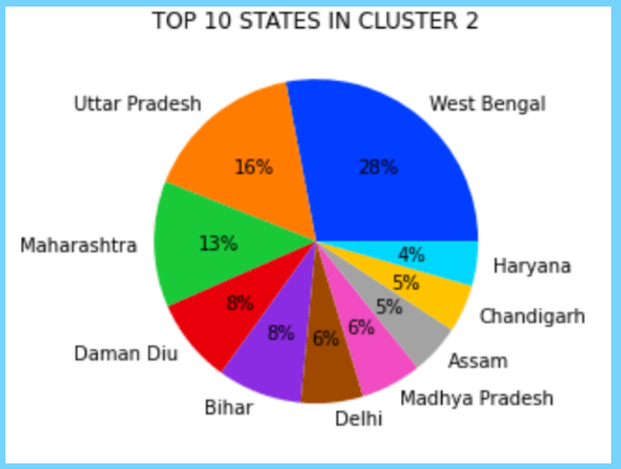


Cluster Summary

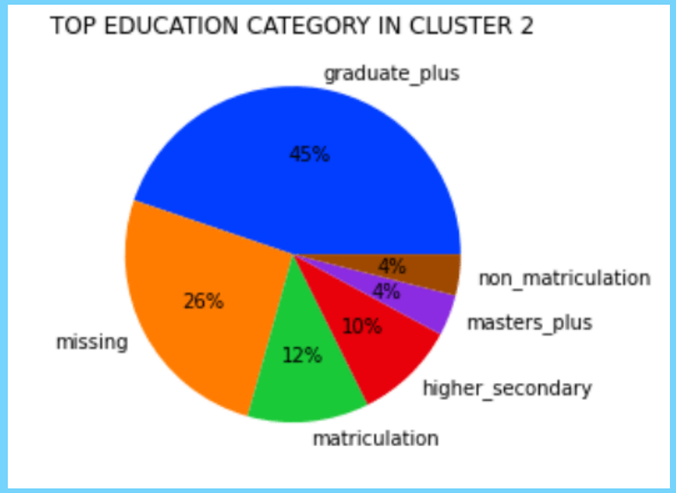
Cluster 3





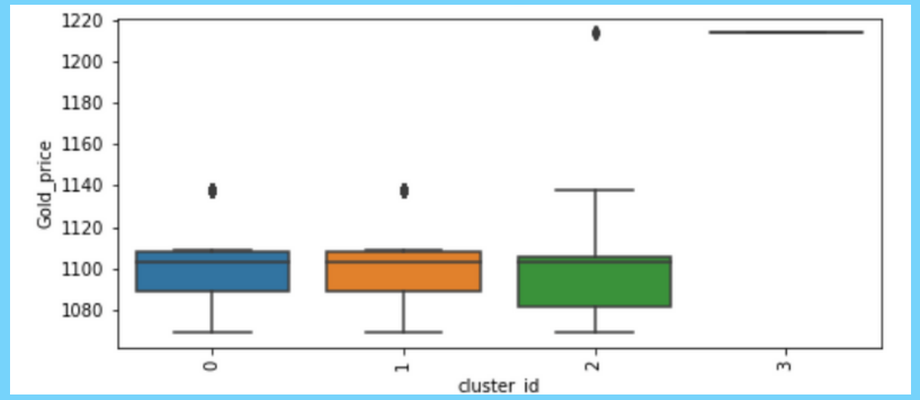


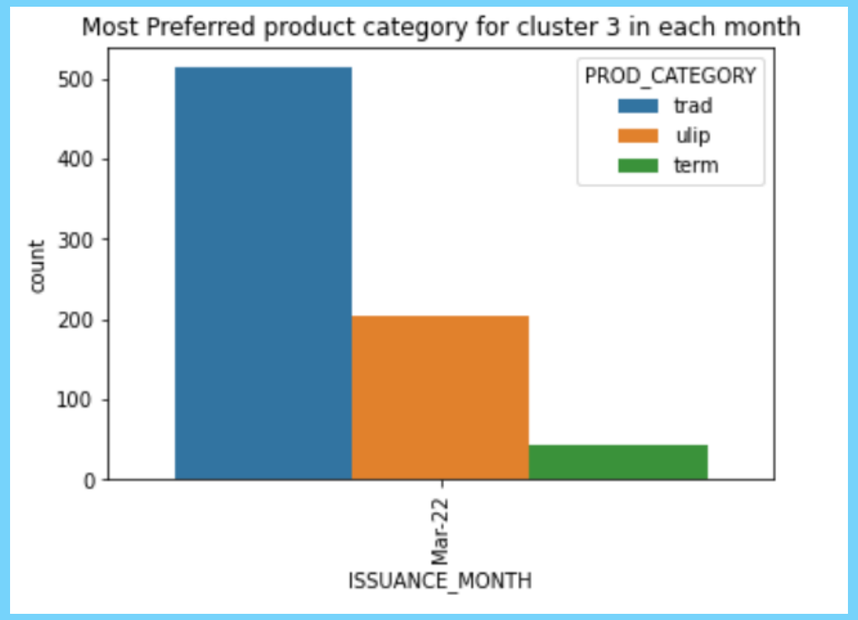
- For this cluster, the ages vary from 40 to 80 and the mean age was around 50.
- Clients from this Cluster are given first preferences to Trad insurance and then Ulip.
- In states like Maharashtra, Chandigarh, and Delhi clients give first preference to the Ulip product category.
- For this cluster clients mostly comes from states like West Bengal, Utter Pradesh, and Maharashtra.



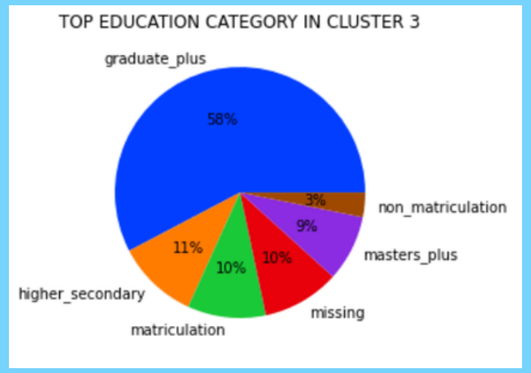
Cluster Summary

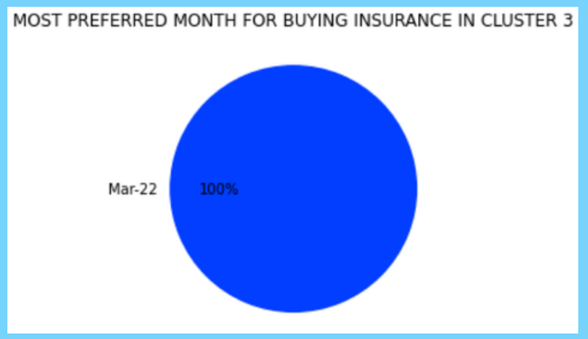
Cluster 4





- For this cluster, the ages vary from 20 to 70 and the mean age was 40.
- Clients from this Cluster are given first preferences to Trad insurance and then Ulip.
- Clients from these clusters mostly come from the medium-income segment category.
- mostly clients are buying insurance when the gold price is high as compared to other clusters.
- Clients are mostly buying insurance in the month of March.
- Most clients are having Education type graduate plus and higher secondary.



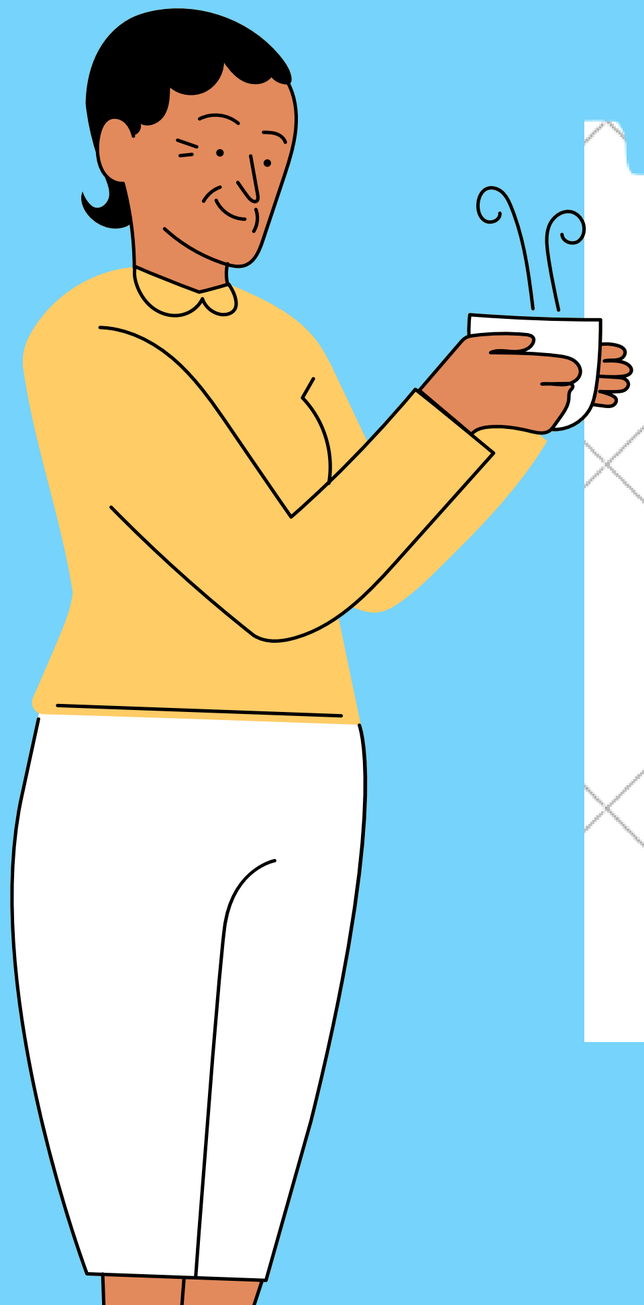


Recommendation

- The features which are mostly affecting insurance plan buying are listed below.
  - From which state the client was from?
  - In which month the client wants to buy the insurance?
  - From which age range the client has belonged?
  - In that particular month, what is the price of gold?
  - what was the educational and occupational background of that client?

- So while predicting we have to consider these features.

- The Gradient Boosting model can be used for deployment. It has the best performance in terms of Accuracy.



THANK YOU!

HAPPY DIWALI