BAJAJ ALLIANZ HACKATHON SUMMARY

by Gaurav Malik

PROBLEM STATEMENT

We are given the following problem statements-

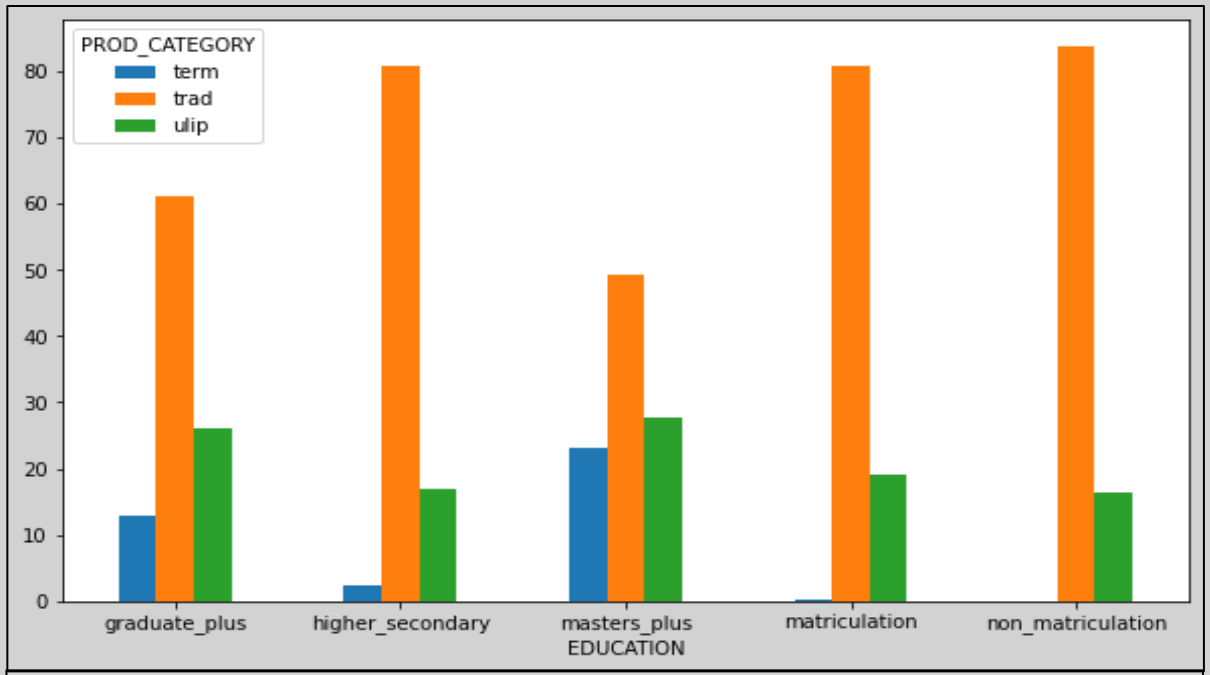
- To understand how people make purchase decision and how they select right insurance policy for themselves.
- To perform accurate population segments using the various features specific to the Indian Market.
- This is classification machine learning problem.
- We must determine the "**prod_catetogry**”
- Initially exploratory data analysis was done, followed by data cleaning, feature engineering, modelling, feature selection.

EXPLORATORY DATA ANALYSIS WALKTHROUGH

- Nan values are present in **age** and **pincode** features.
- Age feature is **skewed**, for this capping is done to limit the skewness.
- There are some **rare labels** in category features, for this if there is any category in particular feature whose frequency is less than 5 percent is been converted to category ‘rare’.
- New feature is created ‘age_category’.
- *Findings which are asked in problem statement are in following slides*.

ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR EDUCATION

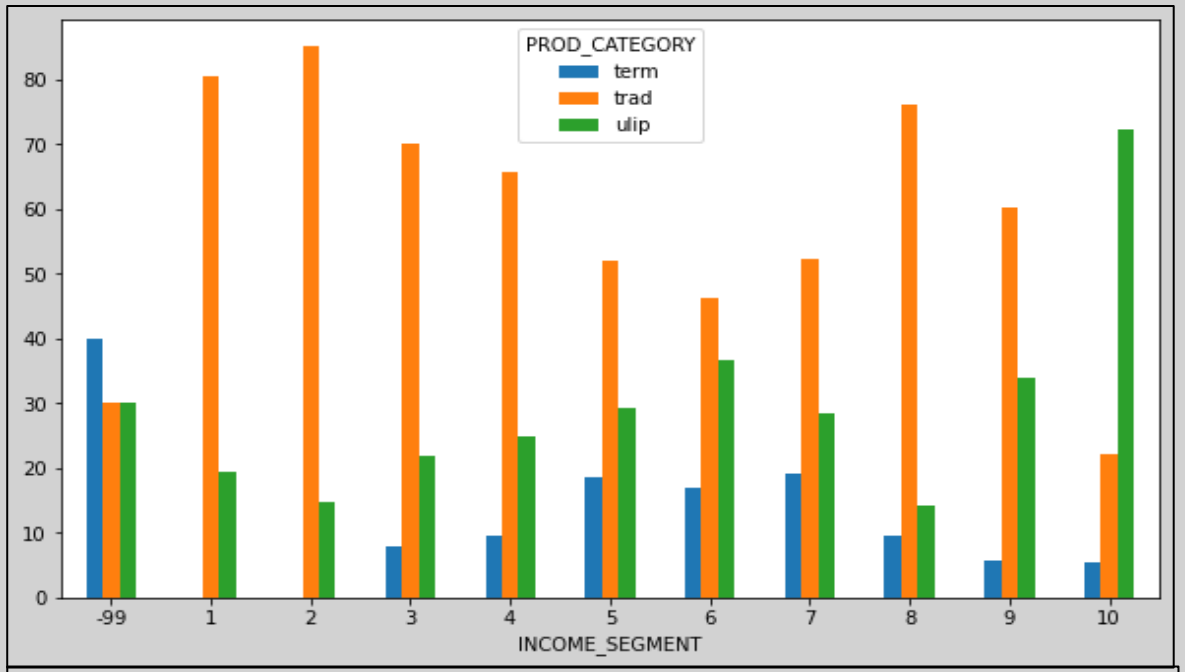
- In **education category**, it seems like **trad and ulip** product category seems to be most preferred.
- It seems like **term** product category preferred by **master_plus** education category.



ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR EDUCATION

ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR INCOME_SEGMENT

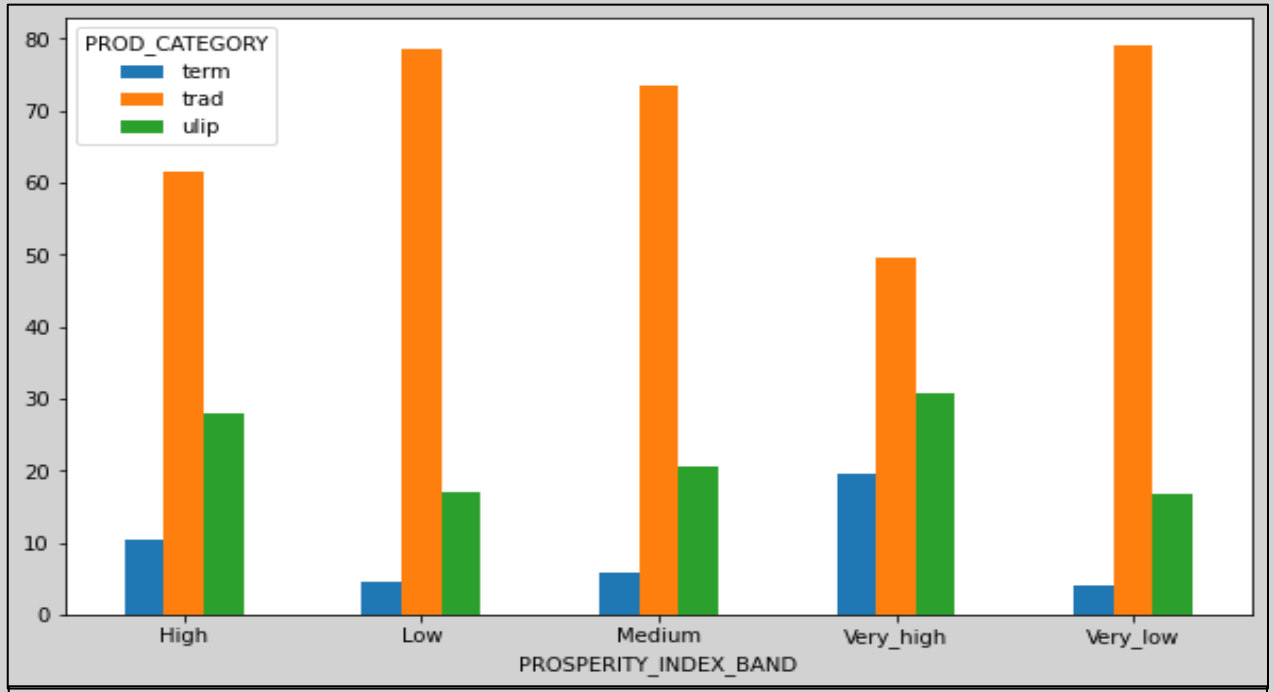
- In **income segment** category, it seems like trad and ulip product category seems to be most preferred.
- It seems like **term** product category preferred by **6** category most.



ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR INCOME_SEGMENT

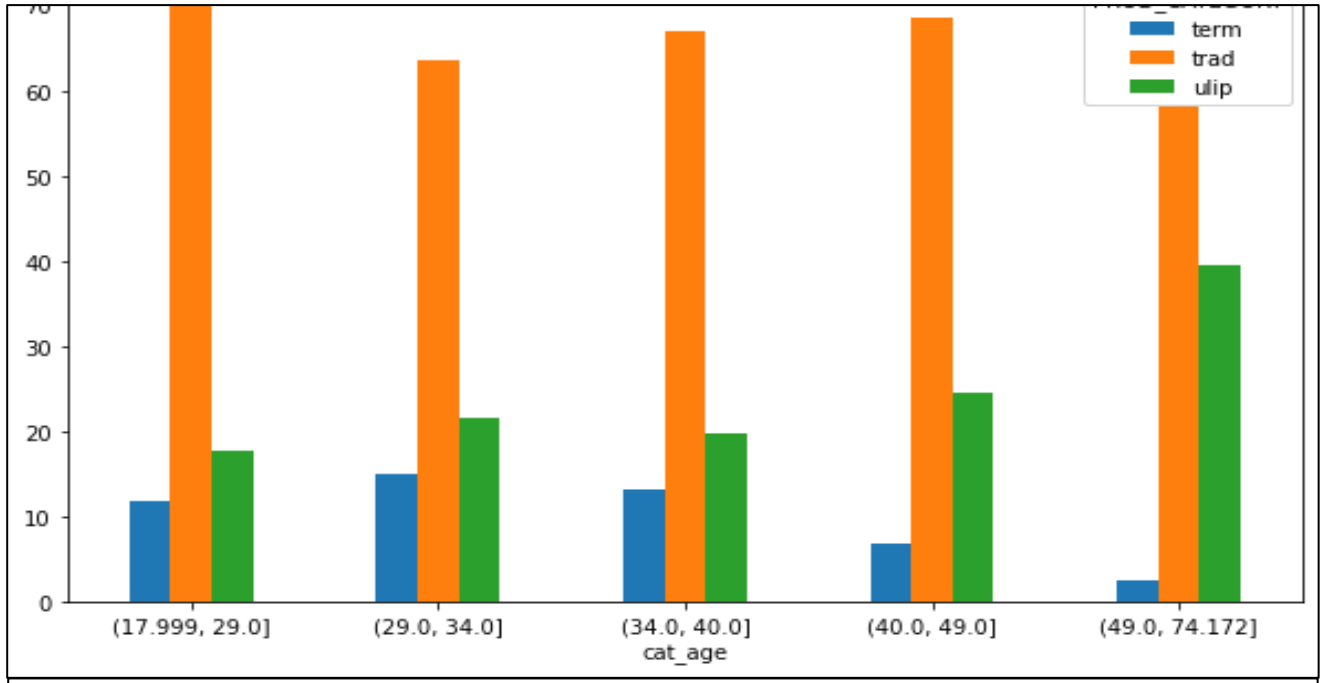
ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR PROSPERITY_INDEX_BAND

- In **prosperity index band** segment category, it seems like trad and ulip product category seems to be most preferred.
- It seems like '**term**' product category preferred by **Very_high** category most.



ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR PROSPERITY_INDEX_BAND

ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR CAT_AGE

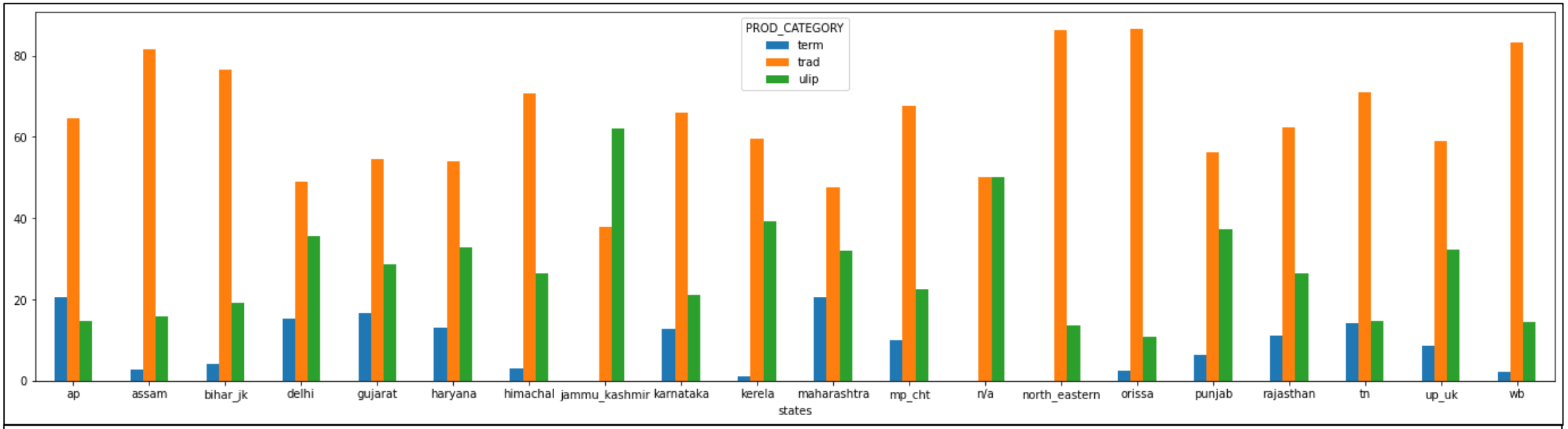


ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR CAT_AGE

- In **cat_age** category, it seems like trad and ulip product category seems to be most preferred.
- It seems like **term** product category preferred by **29-34** year segment category most.

ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR STATE

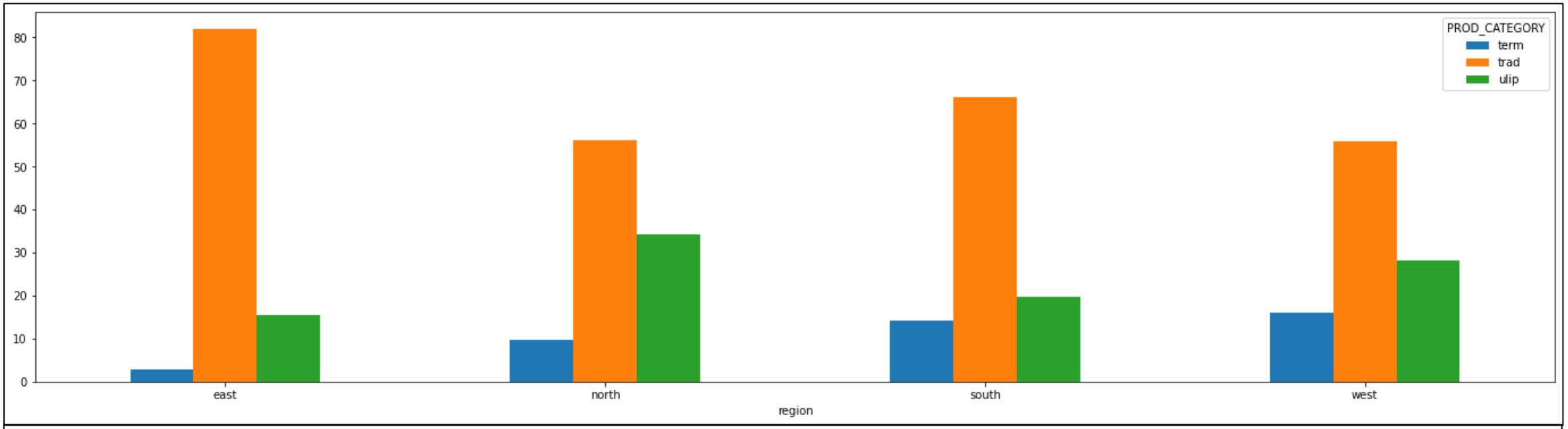
- In **state category**, it seems like trad and ulip product category seems to be most preferred.
- It seems like **term** product category preferred in **Maharashtra and AP** states.



ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR STATE

ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR REGION

- In **state category**, it seems like trad and ulip product category seems to be most preferred.
- It seems like **term** product category preferred in **western** states.



ANALYSIS OF PREFERRED PRODUCT CATEGORY FOR REGION

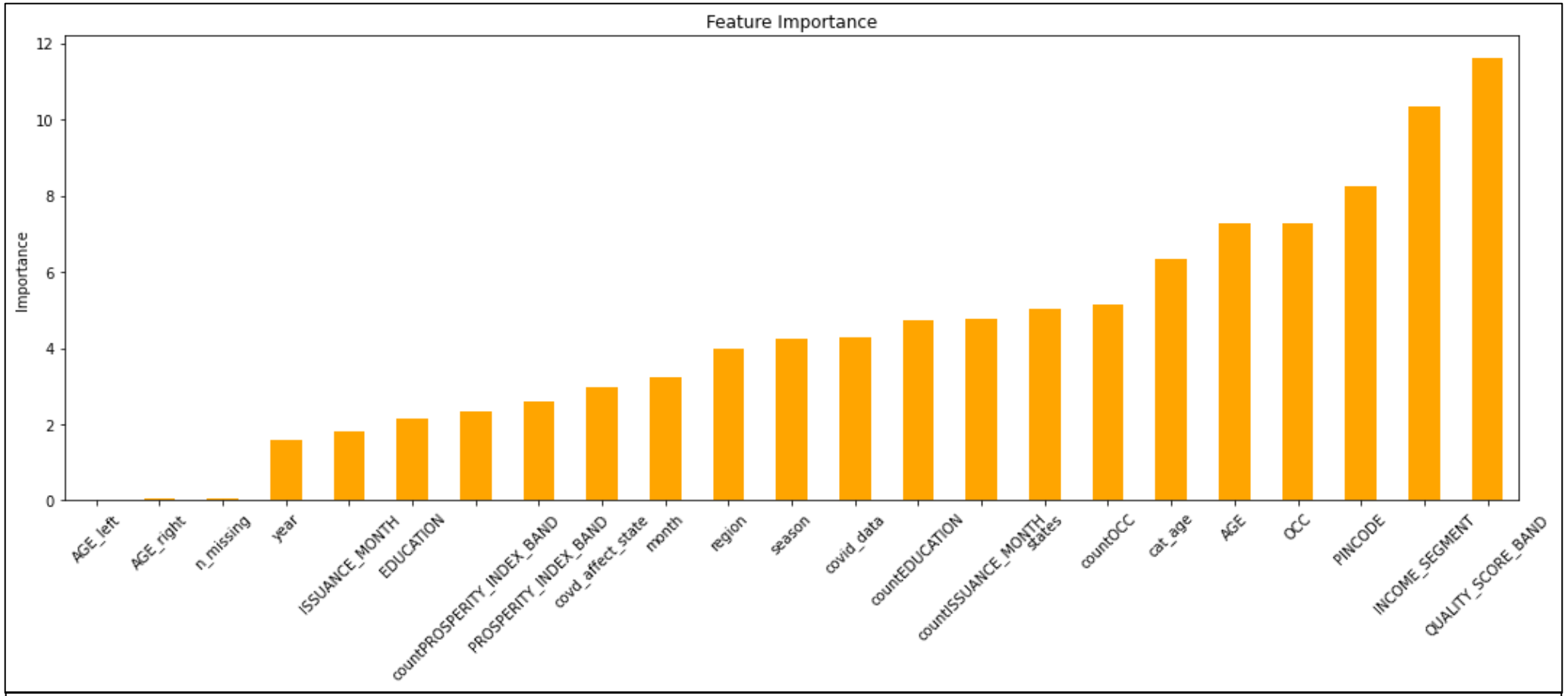
FEATURE ENGINEERING WALKTHROUGH

- New features '**states**' and '**region**' are created on the basis of pincode.
- Other features are also created such as **season, month,** etc.
- Some features are created on the basis of count of category features.
- Things which I tried but didn't worked out-
- I tried adding **Nifty 50 and gold** add but turned out it was giving me less score.
- I also tried **aggregating numerical features** it didn't worked out.

MODELLING WALKTHROUGH

- Initially, I experimented with encoding with various categorical encoders, it turned out **OrdinalEncoder** was giving me good score.
- Secondly I also experimented with imputing with values and not imputing with any values, **KNNImputer** was giving me good score, the 0.72666 score I got from KNNImputer.
- I was also experimenting with different algorithms, initially started with Logistic Regression, subsequently boosting algorithms.
- I decided to go with **catboost**.
- I also tried ensembling and stacking.

FEATURE IMPORTANCE AND WEIGHTAGE



FEATURE IMPORTANCE

THANK YOU