

Creation of segments for Insurance products



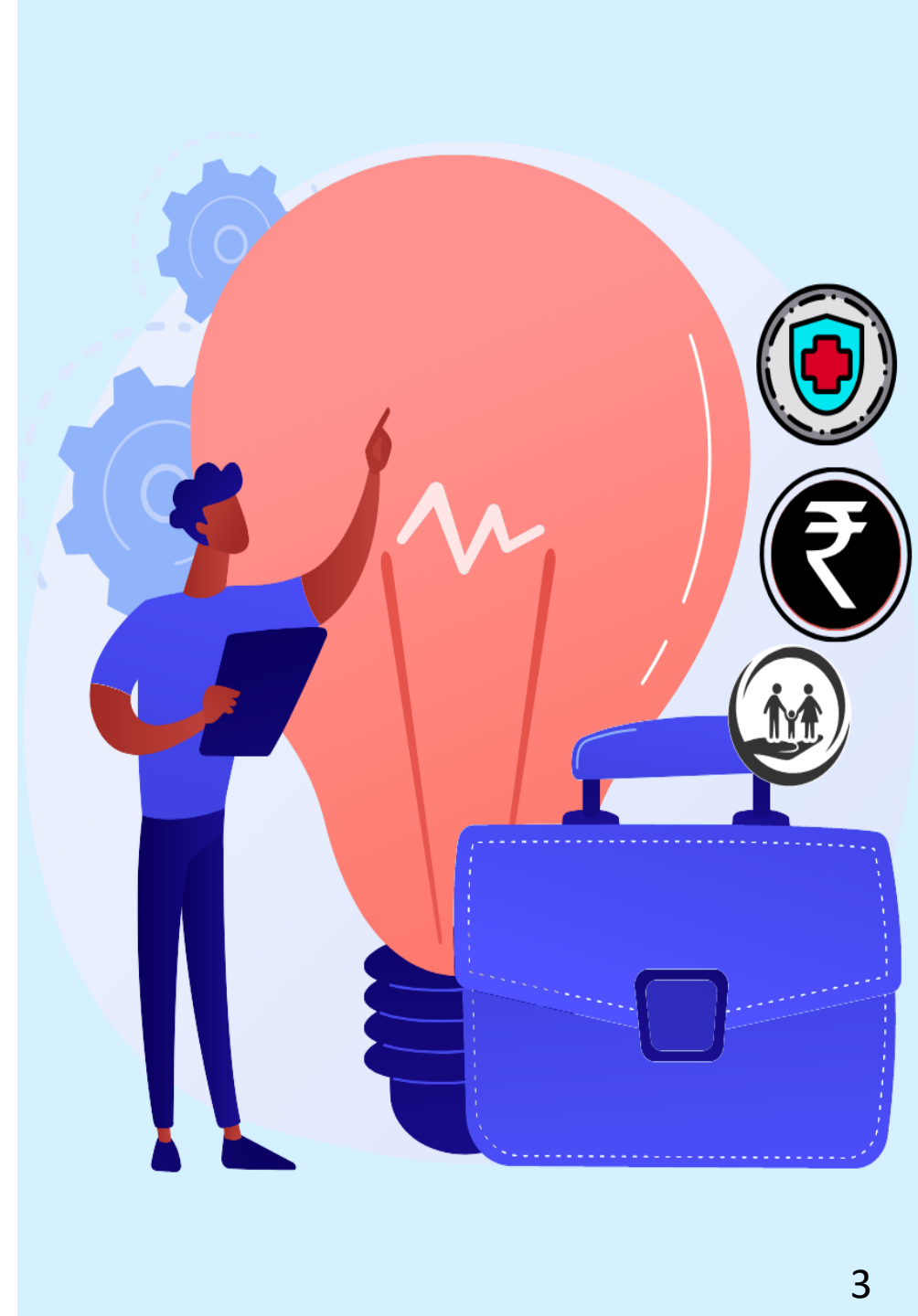
Introduction

- The post-COVID-19 era has witnessed a transformative shift in India's life insurance landscape, marked by a notable increase in penetration rates and a surge in online sales.
- As digital channels become integral to financial product acquisition, consumers are increasingly conducting independent research to inform their insurance decisions.
- This data science project aims to explore and predict the class of insurance chosen by individuals, focusing on **Term Insurance Plans** (Term), **Unit-Linked Insurance Plans** (ULIP), and **Traditional Endowment Plans** (Trad).
- Understanding the factors influencing customer decisions in this dynamic landscape is critical as India embraces digital platforms for insurance purchases.



Problem statement

- Traditional agent influence remains, but many opt for direct online purchases.
- To predict insurance class choices (Term, ULIP, Trad) in this transformed landscape.
- To understand factors influencing insurance choices in a dynamic digital landscape.
- Emphasis on understanding Term, ULIP, and Trad plans in a changing landscape.
- Important to understand customer decision-making and policy selection in this scenario.



Data dictionary

- **ID** - Customer ID
- **AGE** - Age of customer when policy is bought
- **EDUCATION** - Education of customer
- **OCC** - Occupation of customer (professional are Doctors, CA), business is Self-employed
- **PINCODE** - Customer residence pincode
- **PROD_CATEGORY** - Product Type (ulip - Market linked insurance policy, trad - non-market linked policy, term - term product)
- **INCOME_SEGMENT** - Customer Income segment
- **PROSPERITY_INDEX_BAND** - Prosperous location
- **QUALITY_SCORE_BAND** - Customer Quality
- **ISSUANCE_MONTH** - Policy issuance month

4500 rows
13 columns



Included another dataset to check how the **NIFTY50 Stock price** affects the customers behavior in purchasing the insurance.

- **DATE** -
- **NIFTY50_STOCK_PRICE** -
- **BFS_STOCK** -

Approach Taken To Address The Problem Statement



Customer Segmentation

The project aims to create well-defined customer segments based on personas, ensuring diversity and avoiding excessive similarity between segments. This nuanced segmentation will provide a comprehensive understanding of the diverse customer profiles prevalent in the market.



Feature Weightage Assessment

To enhance decision-making insights, the weightage of product features for each identified customer segment will be rigorously assessed. This step is crucial in understanding the varying importance of attributes in the decision-making process, contributing to the development of targeted and effective insurance products.



Recommended System Development

A robust recommender system, whether statistical or algorithmic, will be developed to identify and prioritize attributes that significantly influence customer preferences for specific product features. This system will play a pivotal role in facilitating personalized recommendations tailored to individual customer needs and preferences.



Market Factors & Impact Analysis

The project will undertake a comprehensive analysis to understand how market factors impact different customer segments. This includes assessing the influence of external factors such as interest rates, stock market performance, and gold prices on customer preferences for specific insurance products.

Unsupervised Learning - Clustering



K-Prototype clustering technique is used to segment customers based on similar attributes and policy preferences and below are the observations:

Target Customer Groups:

- Targeting Cluster 0 & Cluster 2 is recommended.
- Target Customers :
 - **Education** – Graduate Plus, Higher Secondary, Matriculation and Masters Plus
 - **Profession** – Business, Private Salaried, Salaried Government and agriculture professionals
 - **Income** – More than 3 to 5
 - **Index** – Medium or High
 - **Quality Score** – Between 4-5

- **Cluster 0:** Recommend a mix of Traditional and ULIP policies tailored for individuals aged 35-49, acknowledging their preference for policies that provide supplementary benefits and investment opportunities.
- **Cluster 1:** Propose a combination of Term and Traditional policies suitable for younger individuals aged 18-34. Emphasize affordable health term policies and underscore the appeal of profitable traditional policies to help them achieve their financial objectives.
- **Cluster 2:** Advocate a blend of ULIP and Traditional policies for individuals aged 50 and above, particularly retirees, highlighting the potential for health insurance coupled with additional earnings from their policies.



Cluster properties (1/2)

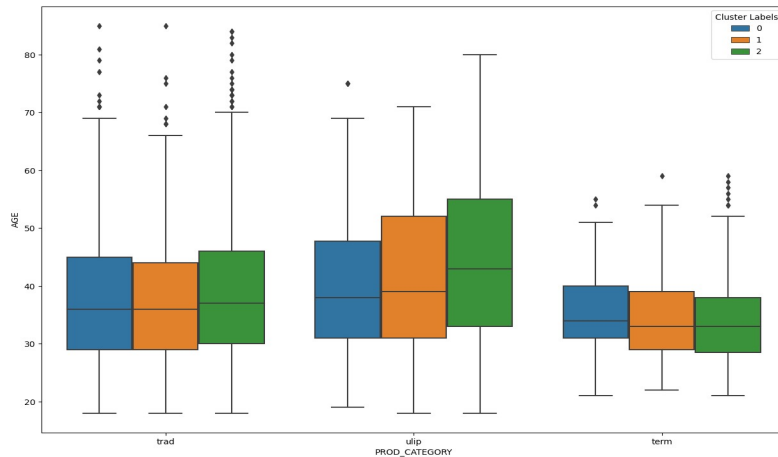
	CLUSTER 0	CLUSTER 1	CLUSTER 2
AGE	35 – 49	18 – 34	49 & Above
EDUCATION	Graduate_plus, Higher Secondary, Matriculation	Master Plus, Graduate Plus	Graduate_plus, Matriculation
OCCUPATION	Business, Salaried_govt, Agriculture	Salaried Private	Business, Retired
PROSPERITY INDEX	Medium & High	Medium & High	Medium & Very High
INCOME	3 – 5	2 – 3	3 – 4



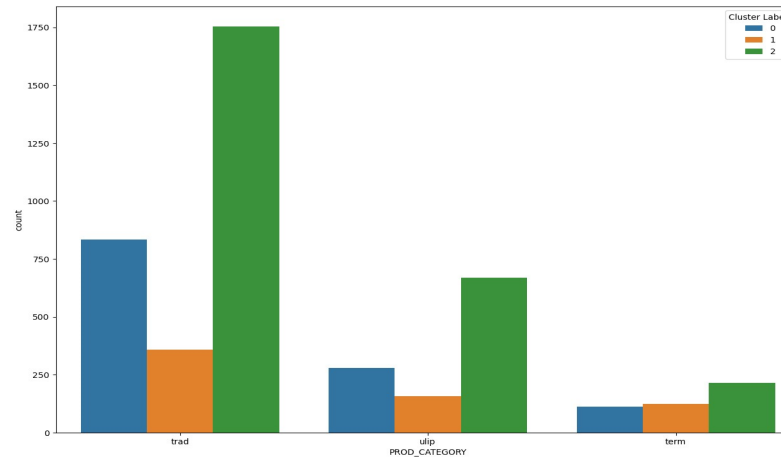
Cluster properties (2/2)

	CLUSTER	PROD_CATEGORY	EDUCATION	OCC	PROSPERITY_INDEX_BAND	AGE	INCOME_SEGMENT	QUALITY_SCORE_BAND
0	0	trad	graduate_plus	business	Medium	41.215583	4.230228	3.826596
1	1	trad	graduate_plus	business	Medium	28.594737	3.478947	4.348947
2	2	trad	graduate_plus	business	High	57.765178	4.564719	3.379152

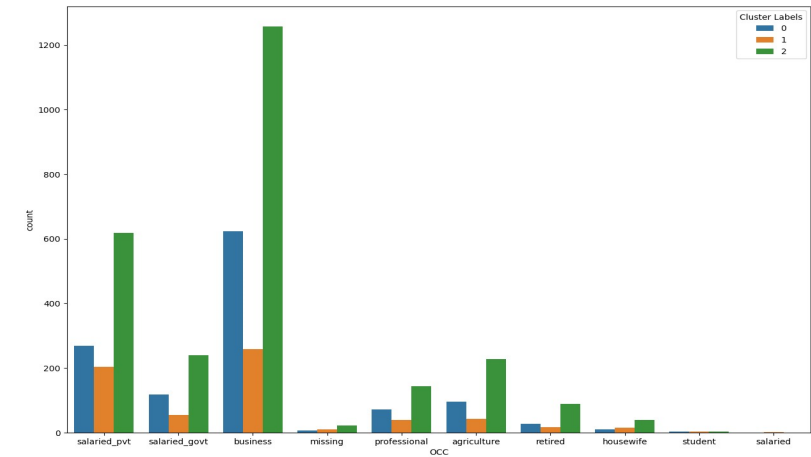
AGE



PRODUCT CATEGORY



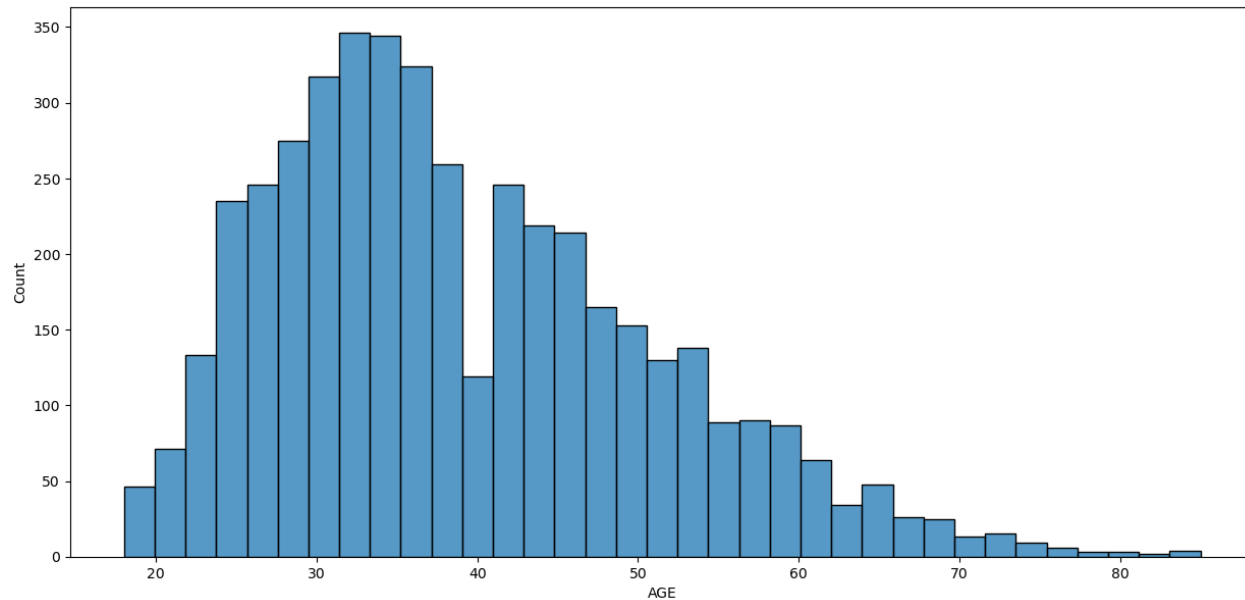
PROFESSION



Feature weightage

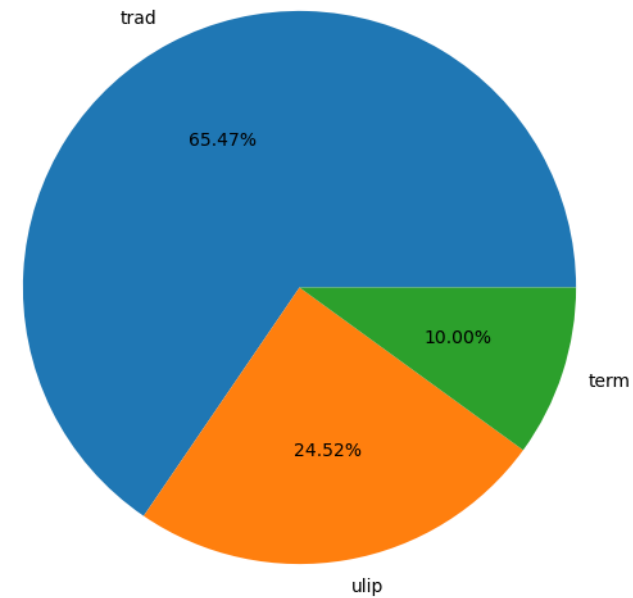


UNIVARIATE ANALYSIS (1/2)



Age-Related Trends:

Individuals within the age range of 30 to 50 demonstrate a heightened proclivity for insurance policy acquisition.



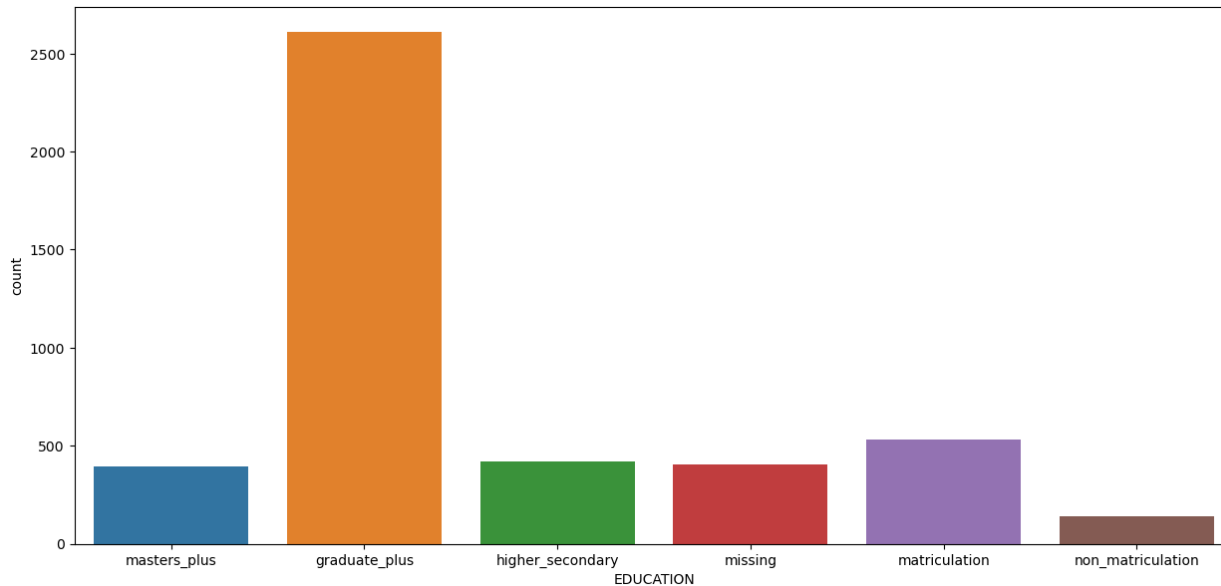
Policy Preference (PROD_CATEGORY):

There is a prevailing inclination among individuals to opt for low-risk insurance policies (Trad) over high-risk (ULIP) or risk-free (Term) alternatives.



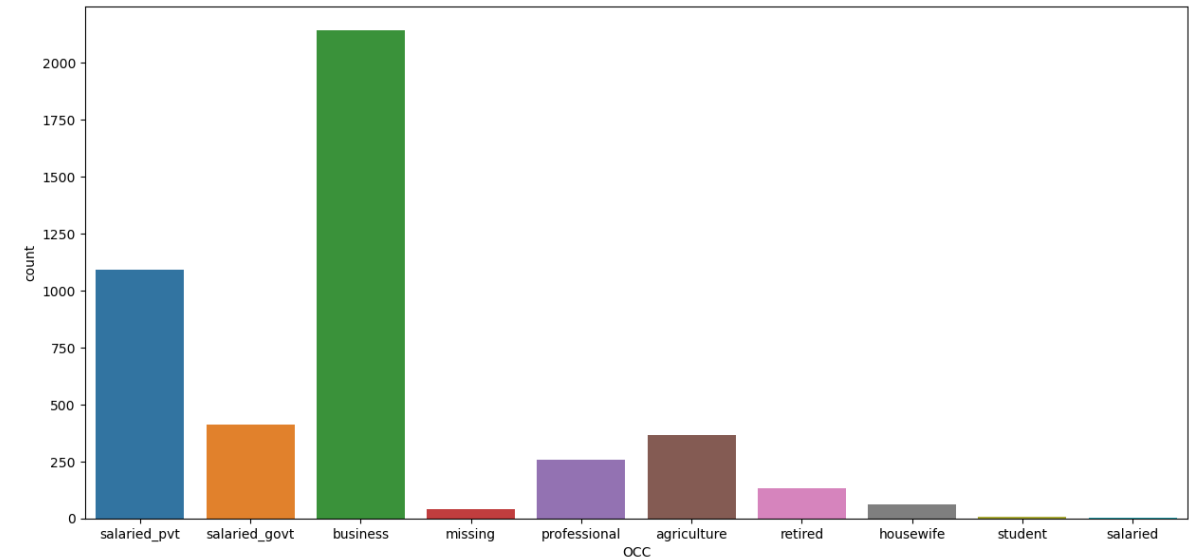
Feature weightage

UNIVARIATE ANALYSIS (2/2)



Educational Background:

Individuals who have completed their education, particularly those with postgraduate degrees, exhibit a higher propensity to engage in the purchase of insurance policies.



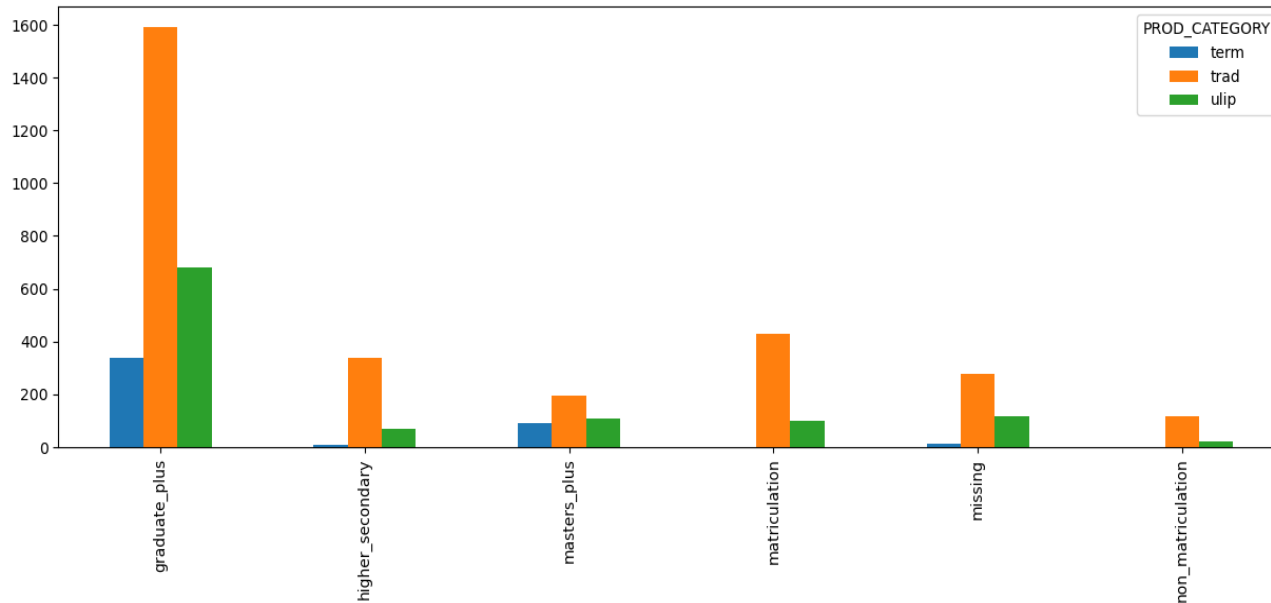
Occupational Influence:

Entrepreneurs and individuals employed in private positions (salaried_pvt) demonstrate a heightened inclination towards procuring insurance policies in comparison to other occupational categories.

Feature weightage

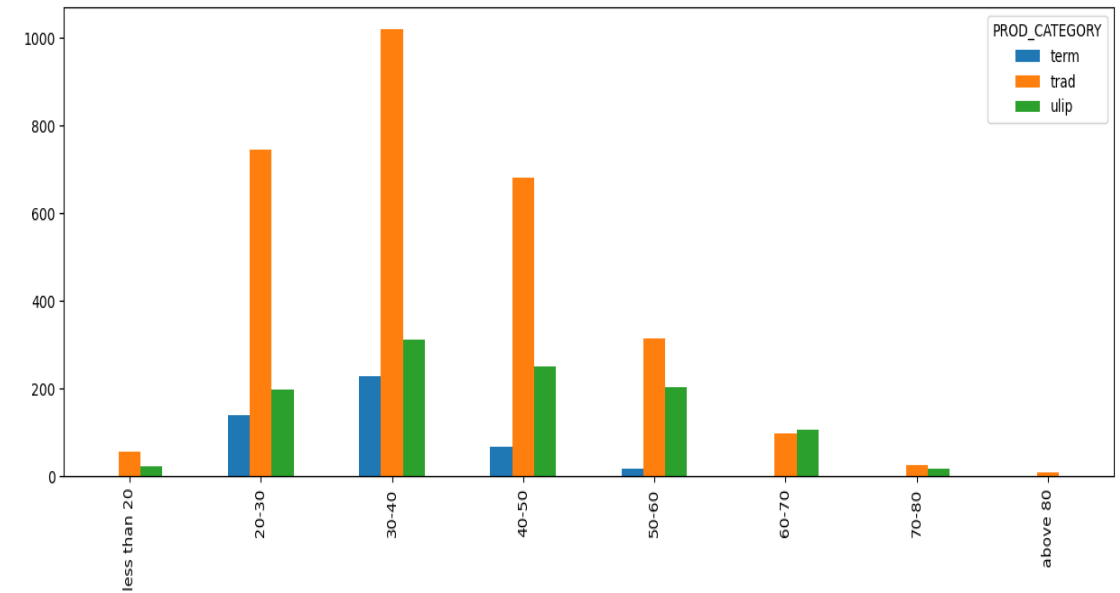


BIVARIATE ANALYSIS (1/2)



Educational Impact on Policies:

Graduates and postgraduates, denoted as graduate plus and masters plus, respectively, exhibit a heightened propensity for acquiring term policies.



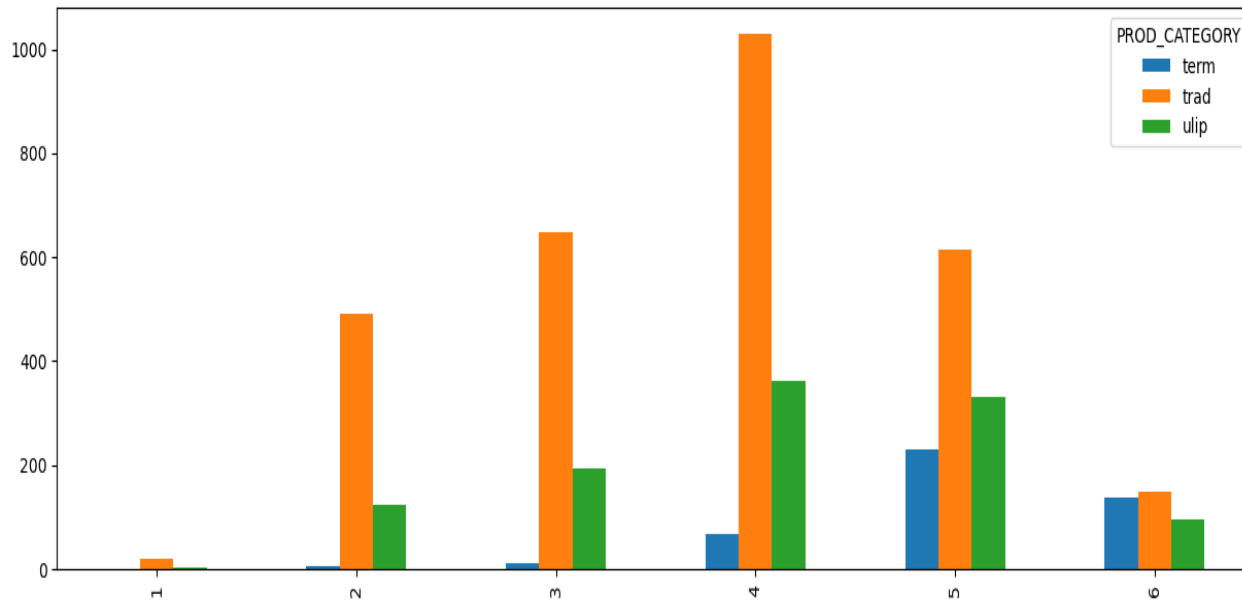
Age impact on Policies:

People aged 20-45 prefer term policies, 40-50 favor both traditional and ULIP policies, and those above 50 prefer ULIP policies.

Feature weightage

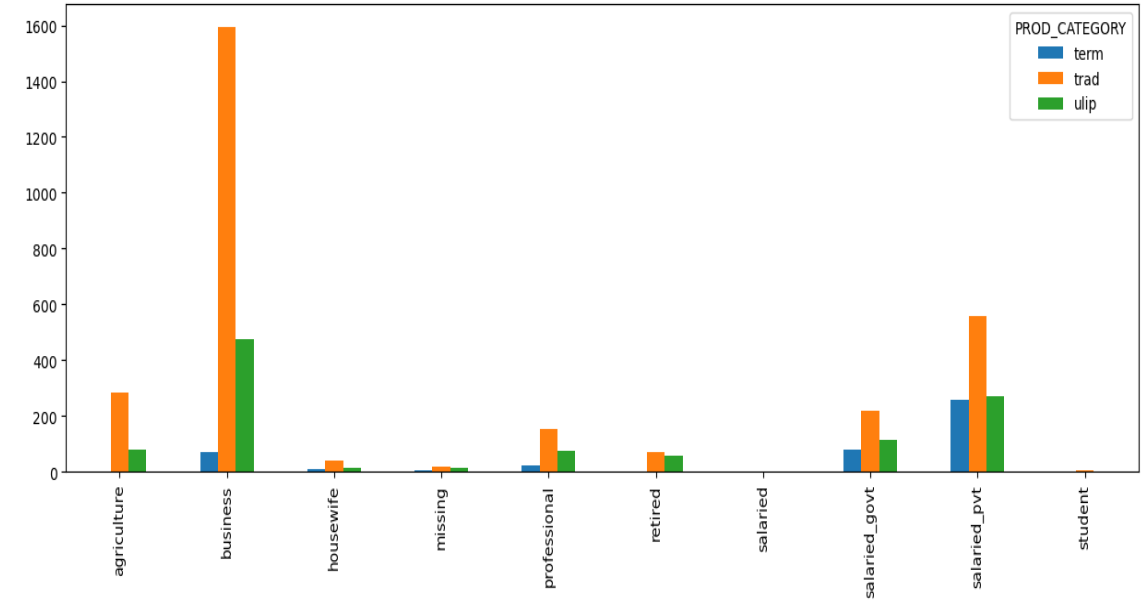


BIVARIATE ANALYSIS (2/2)



Quality Score impact on Policies:

Customers within quality score bands 1 to 4 demonstrate a preference for traditional insurance, while those within bands 5 to 6 exhibit a preference for term insurance.

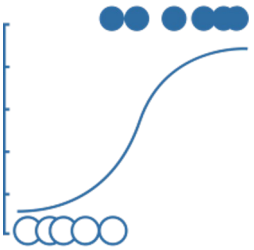


Occupational impact on Policies:

Retirees, farmers, and students favor high-risk (ULIP) and low-risk (Trad) policies.

Salaried individuals, in both private and govt sectors, & housewives exhibit a 100% preference for traditional policies.

Supervised Learning - Classification



Logistic Regression Challenges:

- Struggles with imbalanced multi-class scenarios.
- Biased predictions towards the majority class.



Tree based models:

- **Decision Tree:** Captures complex data relationships, resilient to imbalance.
- **Random Forest:** Ensemble strength for improved accuracy in imbalanced settings.



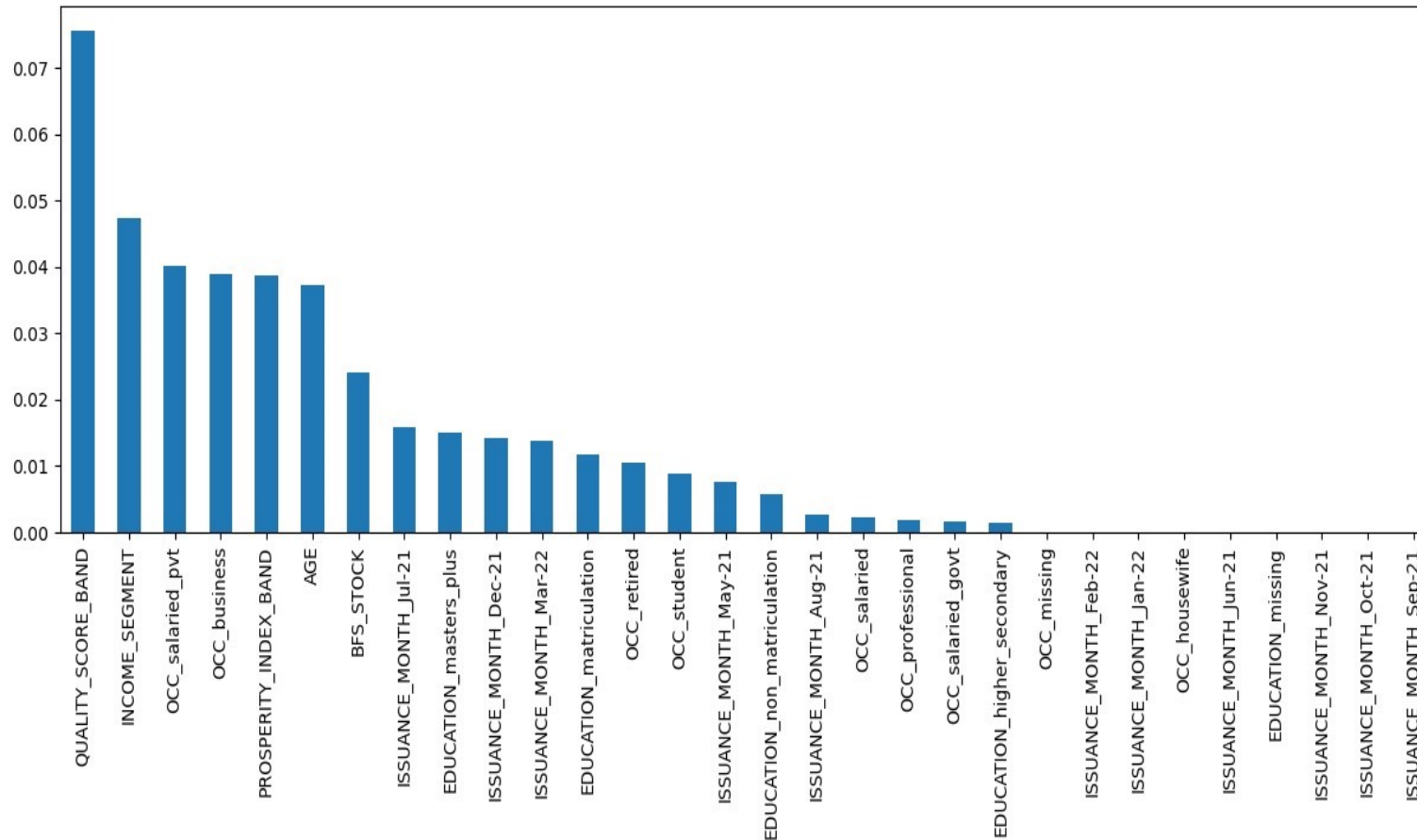
Balanced Metric Emphasis:

- **Precision:** Ensures accurate positive predictions, minimizes false positives.
- **Recall:** Captures true positives, minimizes false negatives.
- **F1 Score:** Harmonizes precision and recall for comprehensive evaluation.

Feature weightage



Base model:-



Feature selection using mutual information classification to improve model efficiency

Mutual Information:

- Definition: Measures the dependence between two variables.
- Use in Feature Selection: Identifies features that carry the most information about the target variable.

Mutual Information Classification:

- Approach: Focuses on selecting features based on their information gain with respect to the target variable.
- Importance: Improves model efficiency by retaining only the most informative features.

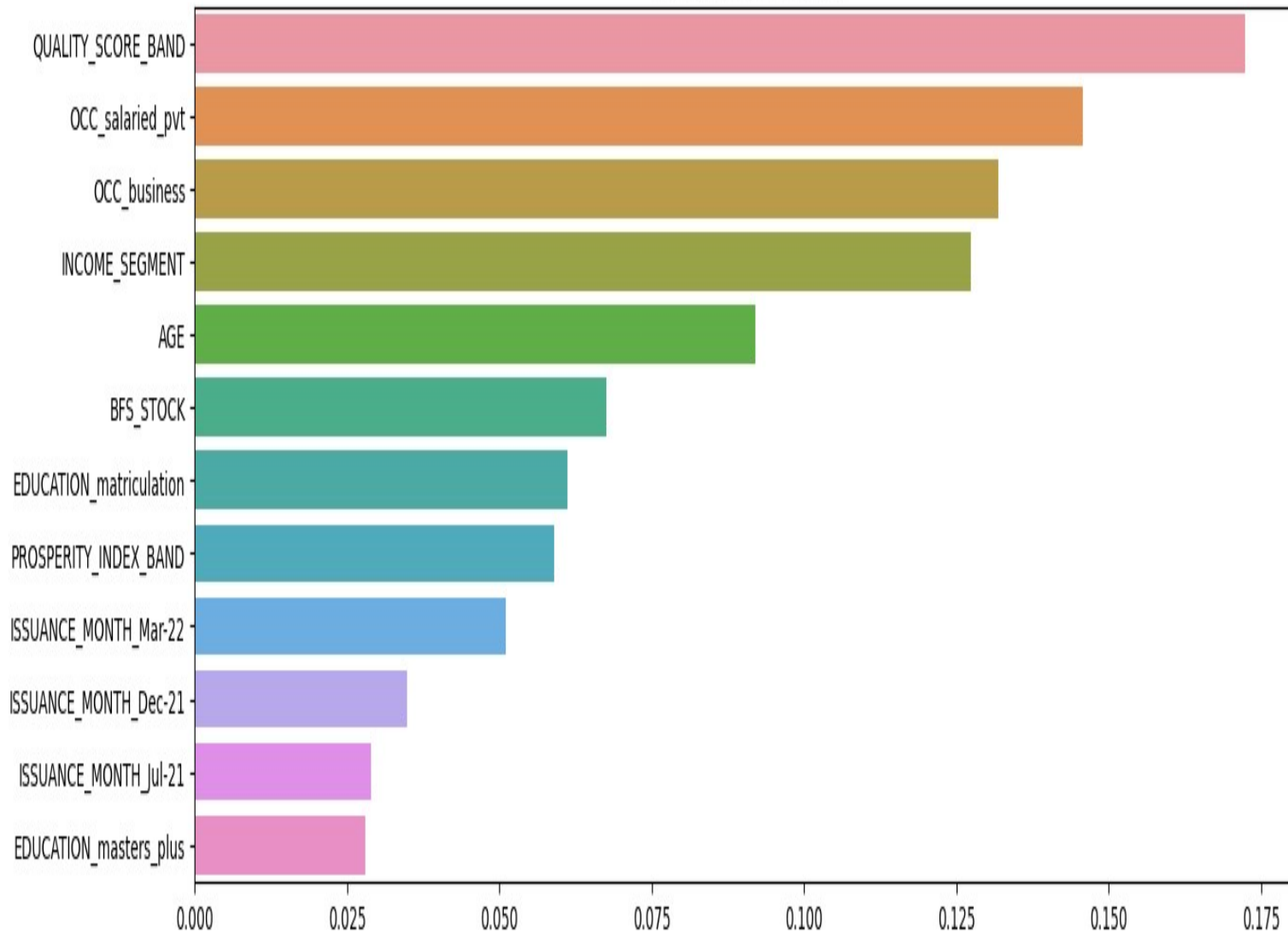
Outcome: A more streamlined set of features that contribute significantly to the classification task, reducing computational complexity and potentially improving model generalization.



Recommender system

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	F1 Score
0	Logistic Regression	0.661372	0.654815	0.661372	0.654815	0.605514	0.562864	0.559399
1	Decision Tree	0.558132	0.543704	0.558132	0.543704	0.667528	0.650458	0.566301
2	Decision Tree Tuned	0.626747	0.566667	0.626747	0.566667	0.690421	0.629862	0.581672
3	Random Forest	0.599111	0.588889	0.599111	0.588889	0.677772	0.669184	0.606946
4	Random Forest Tuned	0.980940	0.626667	0.980940	0.626667	0.981297	0.597364	0.607346
5	AdaBoost with Tuned	0.672173	0.668889	0.672173	0.668889	0.635002	0.630123	0.603695
6	GradientBoost with Tuned	0.686785	0.665926	0.686785	0.665926	0.663827	0.624693	0.600973
7	Stacking Classifier	0.738564	0.648148	0.738564	0.648148	0.763880	0.586521	0.578738
8	SVM Classifier	0.553367	0.562222	0.553367	0.562222	0.541911	0.551360	0.554824
9	XGBoost with Tuned	0.691550	0.674815	0.691550	0.674815	0.679871	0.636748	0.606961
10	Base_Model Logistic Regression	0.664536	0.672222	0.664536	0.672222	0.614567	0.638924	0.606486
11	Base_Model Random Forest	0.607004	0.567778	0.607004	0.567778	0.680424	0.645701	0.586365

Recommender system



AdaBoost Principles:

- Leverages adaptive boosting for imbalanced dataset challenges.
- Assigns weights to weak learners for enhanced effectiveness,

Limitation: Prefers only high Quality data(No outliers)

F1 Score as Evaluation Metric:

- Decision to use F1 score prioritizes balanced precision and recall.
- Crucial for robust assessment in imbalanced class scenarios.

Feature Importance:

- AdaBoost provides feature importance.
- Enhances interpretability and model efficiency

Market factors



		Target Audience	Recommendation Policies
	Educational Impact	Graduates and postgraduates (graduate plus and masters plus).	Term policies are recommended for this educated demographic.
	Occupational Dynamics	Retired individuals, individuals in agriculture, and students.	High-risk (ULIP) and low-risk (Trad) policies are suitable for retirees and those in agriculture, while students may prefer high-risk options.
	Prosperity Index Impact	Regions with very high and high prosperity indices.	Term and ULIP policies are recommended for individuals in prosperous regions.
	Age-related Preferences	20 to 45 years: Term policies. 40 to 50 years: Traditional and ULIP policies. Above 50 years: ULIP policies.	Tailor marketing strategies based on age-related preferences.
	Regional Preferences	Western, southern, and northern regions.	Term and ULIP policies are recommended for these regions, while individuals in the eastern region may prefer traditional policies.
	Income Segment Influences	0 to 4: Traditional insurance. 4 to 8: Term insurance. 8 to 10: All policy types (traditional, ULIP, and term).	Customize product offerings based on income segments.
	Quality Score Associations	Bands 1 to 4: Traditional insurance. Bands 5 to 6: Term insurance.	Align marketing strategies with quality score associations.

References



<https://www.kaggle.com/competitions/allianz-hackathon/overview>



https://in.investing.com/indices/s-p-cnx-nifty-historical-data?interval_sec=monthly



https://in.investing.com/equities/bajaj-finserv-limited-historical-data?end_date=1647282600&interval_sec=monthly&st_date=1615746600

Thank You

Namia
