SEGMENT INSURANCE PRODUCT

Namia

1 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 for 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). Also understanding the factors influencing customer decisions in this dynamic landscape is critical as India embraces digital platforms for insurance purchases.[1].

2 PROBLEM STATEMENT

In the dynamic Indian life insurance market, this project decodes customer decision-making amid agent influence and the shift to online research, offering a uniquely tailored model leveraging relevant data for precision and applicability [2].

- **Customer Segmentation**: Create diverse customer segments, avoiding excessive similarity, for a comprehensive understanding of prevalent profiles.
- **Feature Weightage Assessment**: Rigorously assess product feature weightage for each customer segment, aiding the development of targeted insurance products.
- **Recommender System Development**: Develop a robust recommender system, statistical or algorithmic, to identify attributes influencing customer preferences for personalized product features.
- **Market Factors Impact Analysis**: Undertake a comprehensive analysis of how market factors, including interest rates and stock market performance, impact different customer segments.

3 UNDERSTANDING THE DATA

- This dataset, encompasses **4,500** rows containing details about **customer demographics** and **financial information**. It includes crucial features like Customer ID, Age at policy purchase, Education level, Occupation, Residence Pincode, Income Segment, Prosperity Index Band, Quality Score Band, and Policy Issuance Month. Additionally, I have incorporated **Market factors** such as NIFTY50 Stock Price and BFS Stock .
- **Target Variable (PROD_CATEGORY) :**
  1. **Term Insurance** (Term): Coverage for a specific period, affordable premiums for higher sum assured, death benefit to the nominee.
  2. **Unit-linked Insurance Plan** (ULIP): Combines life insurance and investment, market-linked returns, 5-year lock-in, flexible fund allocation, tax benefits.
  3. **Traditional Endowment Plan** (Trad): Life cover with savings, lump sum on maturity, no daily values published, bonuses accumulate and paid at maturity.

4 PROBLEM-SOLVING APPROACH

- **Customer Segmentation :** Using **K-Prototype** unsupervised learning with three clusters (targeting clusters 0 and cluster 2). Traditional and ULIP policies combined with additional benefits and investment options under Cluster 0 suit people in the age bracket of 35 – 49 years. Cluster 1 proposes a mixture of Termed policies and Traditional ones for people of age group 18-34 who are keen in less expensive Health Term policies, which make profit on these Traditional ones. Cluster 2 recommends mixed use of ULIP and Traditional policies for people over fifty years old, focusing on possibilities of healthcare insurances and more earnings form their policies (pensioners).

- **Machine Learning model** : Tree models were preferred over **Logistic regression** because of product category imbalance. Started with Decision Tree moving towards Ensemble modelling where **F1 scoring** is given importance to give equal importance to Precision and Recall. The best model was further improved through application of **AdaBoosting** on the random forest. **Mutual information classification** were also used for feature selection which selected the top 12 features as well.

5 FINAL MODEL

I chose AdaBoost to be the final model since it is efficient in dealing with skewed data sets and complex situations. Ensemble learning using boosting trains weak learners sequentially and adjusts mistakes to form a strong ensemble model. In particular, AdaBoost emphasizes on a weighted voting scheme for accurate models to have significant influence in the prediction process. Iterative training corrects the mistakes, while the algorithm learns how emphasize difficult examples only.

- **Key Advantages**:
  1. Imbalanced Dataset Handling: AdaBoost performs well for imbalanced data (which is relevant to our insurance classification problem wherein specific classes might be rare).
  2. Versatility: The ability of AdaBoost to work with multiple base learners provides for flexible use in an environment where there are diverse circumstances that may affect the classifications.

- **Limitations**: Outliers should be treated earlier as adaboost is sensitive to outliers.

- **Application in Insurance Classification**:

  It is flexible to misclassification in the new insurance market after COVID and matches the nuances of changing customer tastes. With this in mind, Ada-boost is well suited to dealing with iterative learning and making decisions in circumstances of high intricacy of the nature of insuring factors that influence selection and classification (term, ULIP, Trad).

6 CONCLUSION

The analysis reveals insights: Graduates and postgraduates prefer term policies, while retired individuals, agriculture workers, and students lean towards ULIP and traditional policies. Salaried individuals, especially in private and government sectors, tend to opt for traditional policies. Regions with high prosperity indices prefer term and ULIP policies. The pattern of policy acquisitions starts low at the fiscal year's beginning, peaking in the concluding months. Geographically, the western, southern, and northern regions favor term and ULIP policies, while the eastern region prefers traditional policies. Age-related preferences highlight the 20-45 age group favoring term policies, the 40-50 age range showing a preference for both traditional and ULIP policies, and those above 50 leaning towards ULIP policies. Income segments influence policy preferences, with segments 0-4 favoring traditional insurance, 4-8 showing a preference for term insurance, and segments 8-10 exhibiting preferences across all policy types. Quality score bands 1-4 prefer traditional insurance, while bands 5-6 exhibit a preference for term insurance.

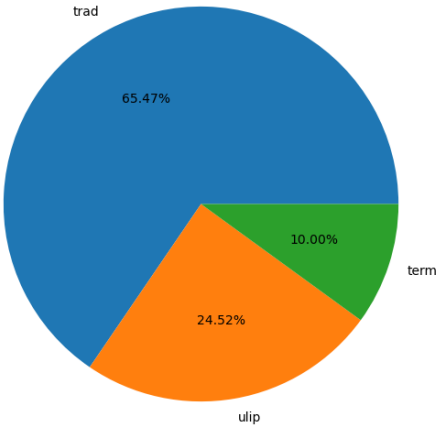

Fig. 1(a) Imbalance in the Product category (Target variable)

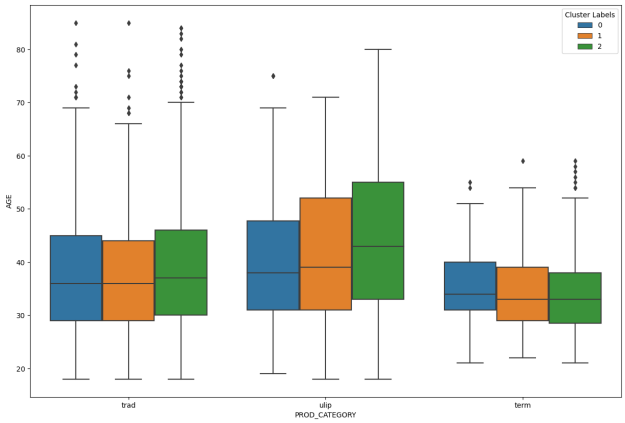

Fig. 1(b) Age distribution across policy types within each cluster

	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 Classfier	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

Figure 2: Classification Model Comparison

7 REFERENCES

1. https://www.kaggle.com/competitions/allianz-hackathon/overview
2. https://in.investing.com/indices/s-p-cnx-nifty-historical-data?interval_sec=monthly
3. https://in.investing.com/equities/bajaj-finserv-limited-historical-data?end_date=1647282600&interval_sec=monthly&st_date=1615746600