Health Insurance Cross-Sell Leads

Impact

In this case study, we focus on Client FinMan, a financial services company offering a range of financial products, including loans, investment funds, and insurance. Our goal is to explore the potential for cross-selling health insurance to existing customers, regardless of their current insurance status with the company. The company uses customer profiles to recommend health insurance when they visit the website. Customers may browse the recommended health insurance policies and proceed to fill out an application form. When customers complete the form, their response is categorized as positive, and they are identified as leads.

Dataset Description and Prep

Dataset

ID	This feature is assigned to each policyholder (or customer) in this dataset, allowing for easy reference and tracking of records.
City_Code	This feature is a code or label representing the city where the policyholder resides or is associated with.
Region_Code	This feature is a code or identifier that denotes the specific region within a city or a broader area where the policyholder is located.
Accomodation_Type	This feature indicates whether the policyholder's accommodation is either 'Owned' (they own their residence) or 'Rented' (they are renting their place of residence).
Reco_Insurance_Type	This feature is a type of insurance recommended to the policyholder, such as 'Individual' or 'Joint'.
Upper_Age	This feature is the maximum age of the policyholder in years.
Lower_Age	This feature is the minimum age of the policyholder, often relevant when multiple individuals are covered under a single policy
Is_Spouse	This feature is certain insurance policies that consider marital status. Indicates whether the policyholder has a spouse ('Yes' or 'No').
Health Indicator	This feature is a categorical variable describing the health condition or risk level of the policyholder.
Holding_Policy_Duration	This feature is the duration for which the policyholder has held an insurance policy with the company.
Holding_Policy_Type	This feature is the type or category of insurance policy that the policyholder currently holds or has held in the past.
Reco_Policy_Cat	This feature is a categorical variable that categorizes the recommended insurance policy into different classes.
Reco_Policy_Premium	This feature is the premium amount that the policyholder is recommended to pay for the insurance policy.
Response	This feature is a binary variable that indicates the policyholder's response to the recommended health insurance policy and is our target feature.

Dataset Here

Data Preprocessing

To enhance our dataset, we must identify potential factors related to customer interest in a premium health insurance plan. We should consider a range of customer demographics and numerical features, including marital status, age range, region, accommodation types, and insurance types. By analyzing these variables, we can determine which customers are more likely to enroll in a premium health insurance plan. This approach also provides valuable insights and analysis, enabling policymakers to design appropriate retention strategies and increase the number of customers with health insurance.

In addition to identifying potential factors that affect customer choices, we must undertake important preprocessing steps:

- **Clean policyholders' columns Holding_Policy_Duration and Holding_Policy_Type:** To clean these columns, we will assign '0' to all their null values. This step allows us to represent non-existing customers by equating their values to '0'.
- **Cleaning Health_Indicator:** To address the Health_Indicator field, whose values are derived from pre-existing metrics based on empirical calculations like BMI, we will assign the missing values using the mode.
- **Resolve Class Imbalance:** We observe a class imbalance in our target feature Response, indicating a higher frequency of '0', which means policyholders do not respond to the recommended health insurance policy. When selecting hyperparameters, we must give greater weight to achieving a balanced dataset.

Exploratory Data Analysis (EDA)

To fully understand the current relationship our explanatory variables have with Response, we first need to visualize the distribution of numeric and categorical variables in our dataset.

- **Numerical Attributes Analysis:** Histograms were plotted for each numerical feature, providing insights into their distribution. Key attributes included Reco_Policy_Premium, Upper_Age, and Lower_Age.
- **Premium Price Analysis:** We examined how the Reco_Policy_Premium relates to customer responses. Boxplots indicated that premium prices had a similar distribution among leads and no-leads, suggesting that the premium price alone might not be a decisive factor in a customer's decision-making process.
- **Premium Price Across Categories:** We analyzed the Reco_Policy_Premium distribution across different categories in Reco_Policy_Cat to understand how policy categories influenced premium pricing and customer responses. Additionally, we compared the Reco_Insurance_Type against the premium prices, revealing that premiums for joint insurance types were generally higher, potentially impacting customer responses.
- **Age-Related Insight:** We compared the age of customers (Upper_Age and Lower_Age) against their responses. The findings suggested a minimal impact of age on response type. However, when health conditions were factored in, certain age-health condition combinations showed a higher likelihood of positive responses.
- **Categorical Attributes Analysis:** We examined categorical attributes to understand their impact on customer responses. The focus was on attributes like City_Code, Accommodation_Type, and Reco_Insurance_Type.

- **Response Rate Across Cities:** Upon analyzing the response rate across different cities, it was observed that some cities had a higher response rate than others, although the overall trend showed more no-leads compared to leads.
- **Accommodation Type Analysis:** Customers renting homes were found to be less likely to purchase insurance if the premium was considered high. This aspect of the analysis highlighted the importance of accommodation status in influencing insurance decisions.
- **Existing vs. New Customers:** We distinguished between existing and new customers using the Holding_Policy_Duration attribute, which suggested that older customers were more likely to accept insurance policy offers compared to newer ones.

Feature Engineering

Creating new features or modifying existing ones can help improve our model's performance. Our feature engineering focuses on transforming numerical variables and encoding categorical ones.

In our numerical feature transformations:

- **Relationship Exploration:** Also known as the relationship between numerical variables, this was assessed using a Spearman correlation heatmap. This analysis revealed notable correlations, particularly between Age and Reco_Policy_Premium.
- **Premium Policy Transformation:** To address the skewness seen in the Reco_Premium_Policy feature, a logarithmic transformation was applied, resulting in a more normalized distribution with fewer outliers. Subsequently, the logarithmically transformed premium prices were categorized into bins of 10 and 50 in both training and testing datasets. This binning process aimed to evenly distribute premium policy prices and improve the model's interpretability.
- **Customer Duration Grouping:** Considering the Holding_Policy_Duration, a binary variable, Long_Term_Cust, was introduced to distinguish long-term customers (with more than 14 days of policy duration) from others. This helps us identify customer loyalty trends.
- **Age Bucketing:** The Upper_Age feature was used to calculate a Mean_Age, which was then categorized into bins of 10 and 50 for both training and testing sets. This helps us segment the customers into different age groups, potentially aiding in targeted marketing strategies to better understand demographics.

In our feature encoding process:

- **Categorization:** We split the dataset into categorical and numerical data types. The categorization of features was based on their nature and relevance to the Response variable.
- **Categorical Features:** Key categorical features included Accommodation_Type, Reco_Insurance_Type, City_Code, and several others. These features were converted to string data types to facilitate their use in predictive models.
- **Numerical Features:** The remaining features were treated as numerical. These included originally numerical features and those transformed or engineered during the analysis. When converted to float data types for consistency, we ensured accurate processing in our models.

- **Target Feature Encoding:** Response was also converted to a float data type, aligning it with the numerical features for modeling purposes.

Machine Learning Model Building

This process involves developing predictive models for binary classification and clustering to understand customer interest in health insurance policies.

Feature Selection and Encoding

- **Target Feature:** The target variable for prediction is `Response` or customer interest in the health policy.
- **Data Splitting:** The dataset is split into training and testing sets, with 30% of data allocated for testing.
- **One-Hot Encoding:** Categorical variables like Accomodation_Type, Reco_Insurance_Type, Is_Spouse, and Cust_Type are transformed into binary columns using OneHotEncoder and SimpleImputer, ensuring that each categorical feature is appropriately encoded for the model.

Logistic Regression

- **Model Choice:** Logistic Regression is used as the primary classifier.
- **Training and Prediction:** The model is trained on the scaled dataset, and predictions are made for both training and testing sets.
- **Model Evaluation:** Various evaluation metrics, including accuracy, precision, recall, F1 score, and ROC-AUC score, are used to assess the model's performance.
- **Observations:** The Logistic Regression model shows moderate accuracy, but its performance in terms of recall and F1 score indicates challenges in handling the imbalanced dataset.

XGBoost

- **Model Choice:** XGBoost, known for handling complex interactions and non-linear relationships, is selected for its robustness.
- **Training and Prediction:** The model is trained on the scaled dataset, with an emphasis on addressing class imbalance.
- **Model Evaluation:** The XGBoost model exhibits a higher accuracy, precision, and F1 score compared to Logistic Regression, indicating better handling of the dataset's complexity and imbalance.

Clustering Model: K-Means

- **Feature Scaling:** Numerical features like Region_Code, Mean_Age, and Reco_Policy_Premium are scaled using StandardScaler.
- **Optimal Clusters Determination:** An inertia plot is used to determine the optimal number of clusters, with a significant drop observed at 3 clusters.
- **Model Application:** K-Means clustering with 3 clusters is applied to the dataset.
- **Cluster Analysis:** The clustering results are visualized to understand the grouping patterns in the data.
- **Insights:** The clustering reveals distinct patterns in customer demographics and policy premiums. For example, a relationship is observed between lower policy premiums and younger age groups.

Conclusion

The binary classification models, Logistic Regression and XGBoost, accurately predict our customer preferences regarding health insurance policies. Logistic Regression, while simpler, struggled with the dataset's imbalance, whereas XGBoost offered improved performance by capturing complex relationships in the data. The K-Means clustering model further enriched the analysis by uncovering inherent groupings in the dataset based on region, age, and policy premiums. These insights are crucial for tailoring health insurance offerings and understanding customer segmentation.