Insurance Company

Machine Learning and Insights

Benedict Laiman - October 2023

Project Structure

01 Introduction

02 Exploratory Data Analysis

03 Binary Classification

04 Recommender System

05 Customer Segmentation

06 Further Insights

Introduction

Project Goals

1. Create a machine learning model to assist in **selecting potential customers**.
2. Develop a **recommendation engine** for new customers who share similar characteristics.
3. **Segment customers** based on their existing policies.
4. Extract **insights** from a given dataset.

Raw Data		
1. ID	6. Insurance Type	11. Policy Type
2. Customer ID	7. Age	12 Policy Category
3. City Code	8. Married	13. Premium Amount
4. Region Code	9. Plan Code	14. Response
5. Accomodation Ownership	10. Policy Duration	

Exploratory Data Analysis

	Dtype	Total_Nan	Nan_Pct	Num_Unique	Example
ID	int64	0	0.0%	50882	[32003, 32285, 2530, 43305, 15714, 23987, 1734...
Customer ID	float64	13	0.03%	18419	[81040.0, 88349.0, 80799.0, 72065.0, 84093.0, ...
City Code	object	0	0.0%	36	[C3, C4, C6, C11, C2, C7, C15, C16]
Region Code	int64	0	0.0%	5316	[3029, 2583, 4479, 267, 4534, 4012, 3721, 329]
Accomodation Ownership	object	0	0.0%	2	[Rented, Owned]
Insurance Type	object	0	0.0%	6	[Individual, Joint, joint, Gabungan, Sendiri, ...
Age	float64	8	0.02%	65	[65.0, 19.0, 24.0, 66.0, 52.0, 56.0, 28.0, 58.0]
Married	object	0	0.0%	2	[No, Yes]
Plan Code	object	11691	22.98%	9	[nan, X2, X3, X1, X6, X4, X5, X7]
Policy Duration	object	20251	39.8%	15	[nan, 14+, 10.0, 4.0, 3.0, 7.0, 2.0, 1.0]
Policy Type	float64	20251	39.8%	4	[nan, 2.0, 3.0, 1.0, 4.0]
Policy Category	int64	0	0.0%	22	[22, 12, 2, 3, 1, 21, 9, 16]
Premium Amount	float64	11	0.02%	28817	[40171.0, 13308.0, 41781.0, 15383.0, 43924.0, ...
Response	int64	0	0.0%	2	[1, 0]

Raw Data

Preprocessing	
1. Dropping 'ID' Column	5. Fixing data types in certain columns
2. Removing missing values rows in several features namely 'Customer ID', 'Age', and 'Premium Amount'	6. Dealing with multiple entries of customers
3. Categorizing variables in 'Insurance Type' column into 'Individual' and 'Joint'	7. Organizing columns into numerical and categorical
4. Dealing with missing values in 'Plan Code,' 'Policy Duration,' and 'Policy Type'	8. Descriptive Statistic Analysis & Feature Engineering

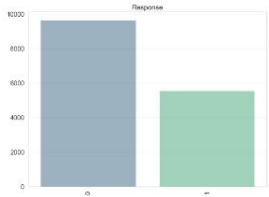
02 - EDA | 01-BenedictLaiman-DS-PP.ipynb

Exploratory Data Analysis

Engineered Features

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15188 entries, 1 to 15203
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             15188 non-null  object
 1   City Code               15188 non-null  object
 2   Accomodation Ownership  15188 non-null  object
 3   Age                     15188 non-null  int64
 4   Married_Insured         15188 non-null  object
 5   Plan Code               15188 non-null  object
 6   Policy Duration         15188 non-null  object
 7   Policy Type             15188 non-null  object
 8   Policy Category         15188 non-null  object
 9   Premium Amount          15188 non-null  float64
 10  Response                15188 non-null  int32
 11  customer_lifetime       15188 non-null  int32
 12  isLoyal                 15188 non-null  object
 13  premium_lifetime        15188 non-null  float64
 14  purchase_frequency      15188 non-null  int32
 15  prodScore               15188 non-null  object
 16  premium_LTYes           15188 non-null  float64
 17  premium_LTNo            15188 non-null  float64
 18  numProdTried            15188 non-null  int32
 19  impression              15188 non-null  float64
dtypes: float64(5), int32(4), int64(1), object(10)
memory usage: 2.2+ MB
```

The numerical features are **not normally** distributed



The response label exhibits an **imbalance**

After implementing various preprocessing techniques that give more emphasis to the class 1 variable, the 'Response' column is ***less imbalanced*** compared to its previous state.

02 - EDA | 01-BenedictLaiman-DS-PP.ipynb

Binary Classification

Assumptions

H0 Customers **do not buy** the insurance product

H1 Customers **buy** the insurance product

False positives and false negatives play **equally important roles** in determining which customers are going to buy the insurance product.

Metric: F1 Score

Features

- Married_Insured
- Plan Code
- Policy Duration
- Policy Type
- Policy Category
- Premium Amount
- premium_lifetime
- impression

Steps

1. Preprocessing

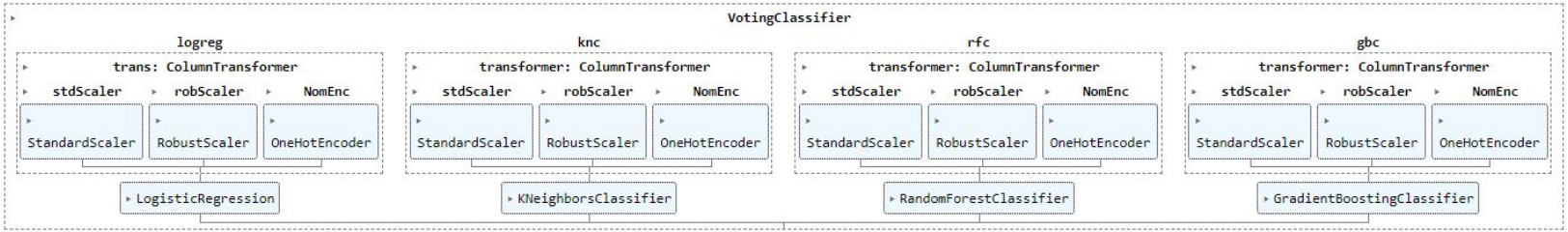
2. Creating Benchmark Classifier Models

3. Fine-tuning the Benchmark Classifier Models

4. Creating Final Classifier Model

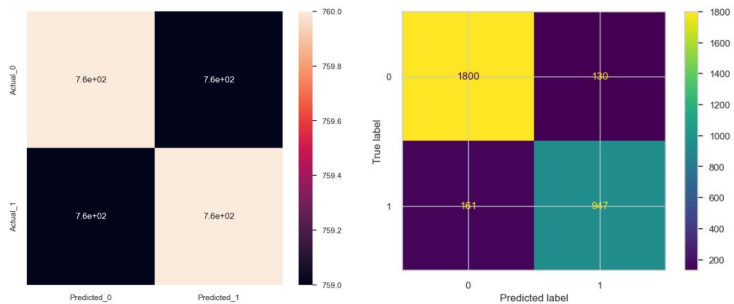
03 - Binary Classification | 03-BenedictLaiman-DS-model-BC.ipynb

Final Classifier Model

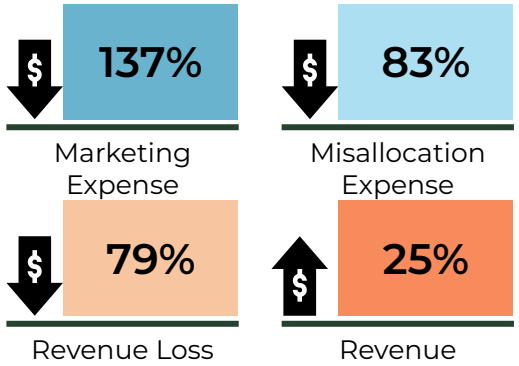


F1

0.90



No Skill | Final Classifier



Overall, the final classifier model **reduces marketing and sales expenses by 220%** and **increases revenue by 103%** compared to the random-guessing model.

03 - Binary Classification | 03-BenedictLaiman-DS-model-BC.ipynb

Recommender System

User-Based Collaborative Filtering

KNN & Cosine Distance	Pearson Correlation
Locating the **closest items** to the target item	Identifying **the most similar customers** to the target customers and providing recommendations based on their Pearson Correlation and ratings

Preprocessing Steps

1. Creating ‘Scoring’ column

2. Creating ‘prodScore’ column

3. Selecting only customers who have tried the insurance product 'more than' once

4. Creating the MN matrix

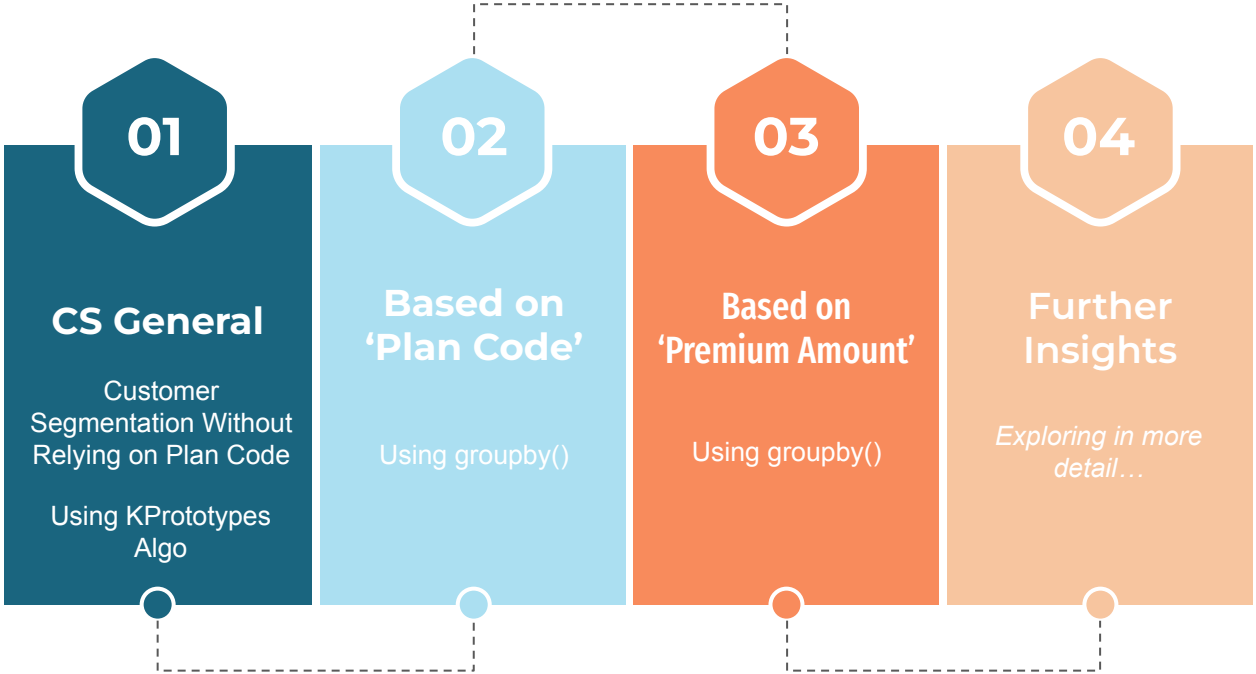
5. Mapping the 'prodScore'

6. Dropping duplicate data

7. Selecting only data with positive ratings

8. Standardizing the customers' ratings

Customer Segmentation



05 - Customer Segmentation | 05-BenedictLaiman-DS-model-CS.ipynb

Customer Segmentation

Heterogen

Homogen

Highest

CS General

	Age	Premium Amount	customer_lifetime	City Code	Married_Insured	Plan Code	Policy Duration	Policy Type	Policy Category	population	percPopulation	meanPremiumAmount	market_size
2	43 - 45	39819 - 40657	0	C2	2	X1	1.0	3	22	1996	36.03	39975.23	79,790,559.08
0	48 - 50	18237 - 18919	0	C1	2	X1	1.0	3	22	2313	41.75	18924.89	43,773,270.57
1	67 - 68	33337 - 34715	35 - 37	C1	2	X1	1.0	3	22	1231	22.22	34235.40	42,143,777.4

Plan Code-based CS

Plan Code	Age	Premium Amount	customer_lifetime	City Code	Married_Insured	Policy Duration	Policy Type	Policy Category	population	percPopulation	meanPremiumAmount	market_size
X1	52 - 54	29222 - 30943	3 - 6	C1	2	1.0	3	22	2378	42.92	30002.41	71,345,730.98
X2	50 - 53	28979 - 31248	2 - 6	C1	2	1.0	3	22	1258	22.71	30056.43	37,810,988.94
X3	50 - 55	28504 - 30974	0 - 5	C1	2	1.0	3	22	755	13.63	29991.51	22,643,590.05
X4	49 - 53	27778 - 31113	2 - 7	C1	2	1.0	3	22	708	12.78	29617.12	20,968,920.96
X5	48 - 56	25013 - 30988	0 - 6	C2	2	1.0	3	22	218	3.94	28638.45	6,243,182.1
X6	55 - 60	28433 - 35390	0 - 10	C2	2	1.0	3	15	155	2.80	30357.46	4,705,406.3
X7	55 - 65	23731 - 38462	0 - 22	C1	2	1.0	3	22	42	0.76	30691.60	1,289,047.2
X8	40 - 63	18307 - 43237	0 - 27	C1	2	1.0	3	15	14	0.25	24671.86	345,406.04
X9	32 - 70	20738 - 43727	0 - 33	C9	2	1.0	3	12	12	0.22	29611.75	355,341.0

05 - Customer Segmentation | 05-BenedictLaiman-DS-model-CS.ipynb

Customer Segmentation

Heterogen

Homogen

Highest

Premium Amount-based CS

	Age	customer_lifetime	Premium Amount	City Code	Married_Insured	Plan Code	Policy Duration	Policy Type	Policy Category	population	percPopulation	meanPremiumAmount	market_size
0	50 - 53	2 - 6	13718 - 14209	C1	2	X1	1.0	3	22	1108	20.0	13948.87	15,455,347.96
1	52 - 54	1 - 6	21756 - 22288	C1	2	X1	1.0	3	22	1108	20.0	21875.01	24,237,511.08
2	52 - 55	3 - 7	29665 - 30101	C1	2	X1	1.0	3	22	1108	20.0	29885.14	33,112,735.12
3	51 - 54	1 - 6	37729 - 38152	C1	2	X1	1.0	3	22	1108	20.0	37977.75	42,079,347.0
4	51 - 55	1 - 5	45530 - 46221	C1	2	X1	1.0	3	22	1108	20.0	45868.84	50,822,674.72

Current Trend Findings

Notable Difference

'Age', 'customer_lifetime', 'Premium Amount' features

Same Traits Exhibition

In most types of customer segmentation

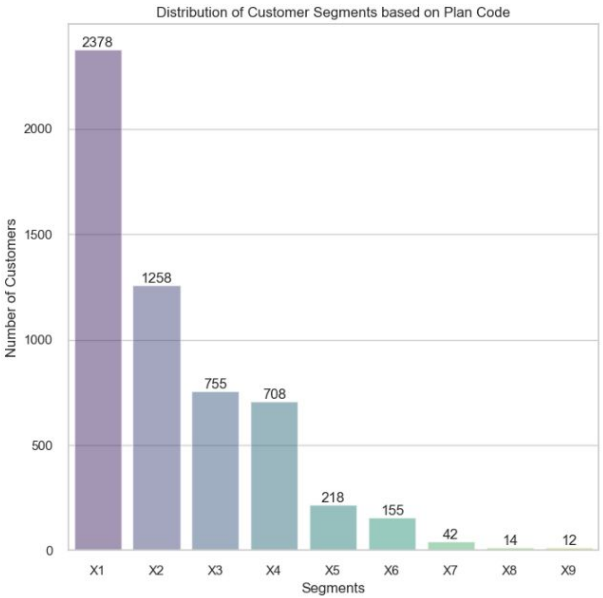
'Married_Insured,' 'Policy Duration', 'Policy Type,' 'Policy Category', and/or 'Plan Code' features

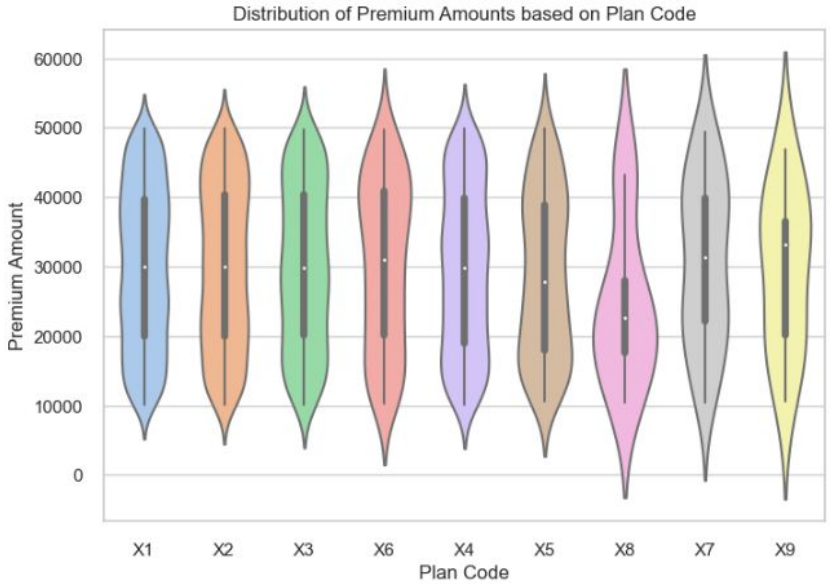
Marketing Strategies

4Ps Product, Price, Place, Promotion

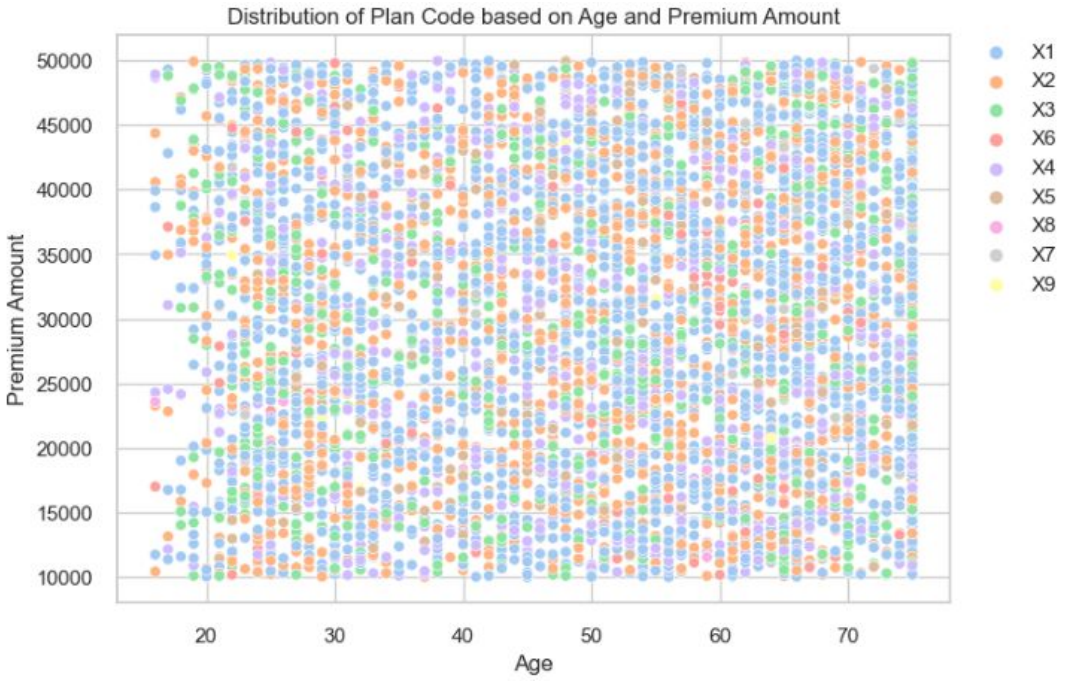
05 - Customer Segmentation | 05-BenedictLaiman-DS-model-CS.ipynb

Further Insights



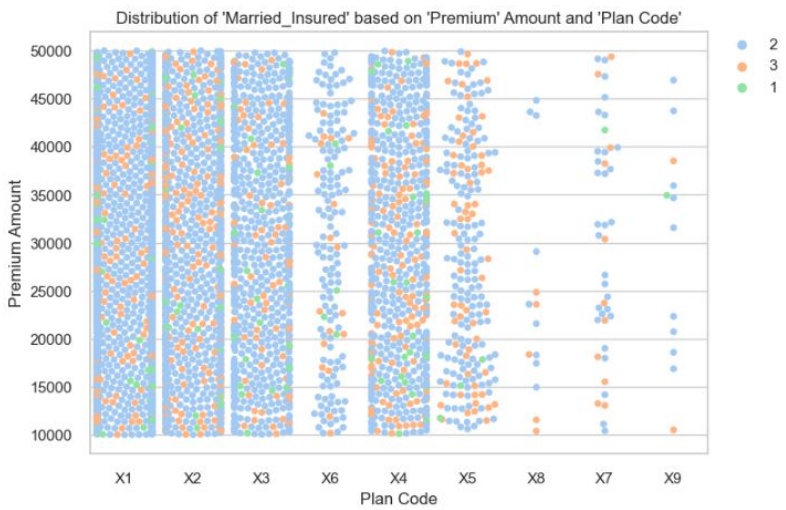


Further Insights



06 - Further Insights | 05-BenedictLaiman-DS-model-CS.ipynb

Further Insights



Further Insights

Conclusion

PABinned	plan_code	Age	customer_lifetime	Premium Amount	City Code	Married_Insured	Policy Duration	Policy Type	Policy Category	population	percPopulation	meanPremiumAmount	market_size	market_sizeFormatted
0.0	X1	49.0 - 54.0	1.0 - 7.0	13566.0 - 14326.0	C1	2	1.0	3	22	452	8.16	13899.83	6282723.16	6,282,723.16
	X2	49.0 - 55.0	0.0 - 12.0	13270.0 - 14319.0	C2	2	1.0	3	22	244	4.40	13904.46	3392688.24	3,392,688.24
	X3	45.0 - 60.0	0.0 - 13.0	13926.0 - 14919.0	C1	2	1.0	3	22	159	2.87	14199.27	2257683.93	2,257,683.93
	X4	45.0 - 55.0	1.0 - 11.0	13500.0 - 14334.0	C1	2	1.0	3	22	151	2.73	13862.88	2093294.88	2,093,294.88
	X5	45.0 - 58.0	0.0 - 5.0	13284.0 - 15391.0	C1	2	1.0	3	22	55	0.99	14335.09	788429.95	788,429.95
	X6	56.0 - 67.0	0.0 - 17.0	12484.0 - 15361.0	[C1, C2]	2	1.0	3	15	35	0.63	13746.57	481129.95	481,129.95
	X7	57.5	13.5	13169.5	C2	[2, 3]	1.0	3	22	6	0.11	12936.50	77619.00	77,619.0
	X8	50.5	11.5	13261.0	[C1, C2, C28, C7]	[2, 3]	1.0	3	15	4	0.07	13593.25	54373.00	54,373.0
	X9	38.5	0.0	13700.5	[C2, C3]	[2, 3]	[11.0, 9.0]	[3, 4]	[13, 2]	2	0.04	13700.50	27401.00	27,401.0
1.0	X1	52.0 - 56.0	1.0 - 8.0	21706.0 - 22439.0	C1	2	1.0	3	22	481	8.68	21882.07	10525275.67	10,525,275.67
	X2	47.0 - 53.0	0.0 - 6.0	21231.0 - 22407.0	C1	2	1.0	3	22	259	4.68	21786.22	5642630.98	5,642,630.98
	X3	49.0 -	0.0 - 8.0	21355.0 -	C1	2	1.0	3	22	140	2.53	22176.67	3104733.80	3,104,733.8

'Plan Code' and 'Premium Amount' that has been quantile-binned

Takeaways

- Swarm plots aid in visualizing data
- Refined segmentation benefits marketing
- Prioritize 'market_size' and tailor strategies

Thank you!