WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n1.14
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research On Purchasing Behavior and Marketing Strategy Optimization of Commercial Medical Insurance Based on Unsupervised Learning Algorithms

Tianyi Ren

School of Economics and Management, Tianjin Polytechnic (Tiangong) University, Tianjin, China
rentianyi46@gmail.com

ABSTRACT

Being a marketer in today's world is not an easy job, especially in the insurance industry, where many factors affect whether customers buy insurance and how much they buy. This field also has many research topics and areas worth exploring. In this article, the data used is real insurance sales data and used unsupervised learning methods in machine learning to clean, mine and analyze the data. First, using clustering methods to divide people who buy insurance into three categories based on weight and insurance expenses. Then using association rule learning methods to analyze the recurring features among people who buy insurance. Finally, based on these results, many policies and suggestions can be given from the perspective of insurance companies and sellers.

KEYWORDS

Unsupervised learning methods; Machine learning; Policies and suggestions; Insurance companies and sellers

1. INTRODUCTION

To understand the current state and challenges of health insurance, it's essential to recognize several key issues affecting the system:

High costs and Accessibility: The high cost of healthcare in the United States is a significant barrier, leading many individuals to skip or delay necessary medical services. Approximately one in four adults have reported postponing or forgoing care due to cost, with this issue disproportionately affecting women, older adults, and uninsured individuals. The financial burden also extends to prescription drugs, with many adults finding it difficult to afford necessary medications, leading to skipped doses or unfilled prescriptions (KFF) (Health System Tracker).

Health Spending: In 2021, U.S. health spending reached nearly $4.3 trillion, accounting for 18.3% of the GDP. This expenditure is more than double the average of other high-income countries. Despite this high spending, out-of-pocket costs continue to rise, putting additional financial pressure on individuals. These costs include expenses for inpatient care, emergency visits, and routine services, which can quickly accumulate, especially for those with serious health conditions (Health System Tracker).

Insurance Coverage and Inequality: Despite the availability of health insurance, there are significant disparities in coverage and access to care. Individuals with lower incomes, people of color, and those living in states that have not expanded Medicaid are more likely to experience difficulties in accessing

healthcare. Moreover, even those with insurance often face substantial out-of-pocket costs, which can lead to medical debt and further financial strain (KFF) (Health System Tracker).

Impact of the COVID-19 Pandemic: The pandemic exacerbated existing challenges in the healthcare system, highlighting the need for improved public health infrastructure and expanded insurance coverage. Many people delayed or avoided seeking care due to pandemic-related barriers, and while telehealth helped bridge some gaps, the utilization of healthcare services has not fully returned to pre-pandemic levels (Health System Tracker).

These challenges underscore the importance of analyzing health insurance purchasing behaviors using machine learning algorithms. Such analysis can identify patterns and trends that inform policy decisions aimed at improving affordability, access, and overall health outcomes. Understanding these behaviors can lead to more tailored insurance products and services that meet the diverse needs of the population, ultimately contributing to a more equitable and efficient healthcare system.

Not only in the United States, but many countries in the world are facing similar problems. Therefore, the analysis of commercial health insurance purchasing behavior, especially quantitative analysis, is necessary and urgent.

As an efficient and practical tool, machine learning is widely used in various fields around the world to solve various problems. This article will use models of unsupervised learning to analyze customer purchasing behavior in commercial medical insurance, discover hidden connections between data, and give strategic recommendations from the perspective of insurance companies. In writing this article, the research results and inspirations of predecessors are very important. The following is a literature review:

In the article *Predicting Credit Card Transaction Fraud Using Machine Learning Algorithms* by Jiaxin Gao, Zirui Zhou, Jiangshan Ai, Bingxin Xia, and Stephen Coggeshall, the authors used a variety of machine learning models and compared their accuracy and differences. This provided me with many ideas for analyzing problems and helped me gain a certain understanding of the various applications of machine learning algorithms.

In Sriramakrishnan Chandrasekaran, Abhishek Kumar's article *A Clustering Approach for Customer Billing Prediction in Mall: A Machine Learning Mechanism*, the authors used clustering methods to focus on whether customers would consume in the mall. They used hierarchical clustering and K-mean algorithms for various characteristics of consumers (gender, age, income, etc.), and used confusion matrices to find the most accurate method. This deepened my understanding of the K-mean clustering algorithm. I also gained a lot of insights into unsupervised learning.

In the article *A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims* by Hojin Moon, Yuan Pu, and Cesarina Ceglia, the authors used algorithms such as logistic regression and random forest to avoid automobile insurance fraud. Among many customer characteristics, they found the factors that are most relevant to whether insurance fraud is committed.

After deeply inspired by my predecessors, the purpose of this article determined to be establish a model that can be applied to insurance sales more efficiently. The main problem solved in this article are:(1) Use clustering algorithms to build a model that divides customers into different groups based on different characteristics, making it easier for managers and sales staff to conduct targeted sales and improve service quality and efficiency. (2) used association rule learning methods to analyze the recurring features among people who buy insurance.

2. MATERIALS AND METHOD

2.1. Research Steps

First, we collect data. The data used in this paper comes from a public dataset, with a total of 25,000 samples; Second, Perform data cleaning; Third, Identify data features and select models; At last, Build, train and test machine learning models.

2.2. Description of the Proposed System

2.2.1. Data collection

In this study, I found real-world insurance company sales data from the Internet, which included 25000 records and 24 variables, It contains basic demographic Information such as gender, age, education level, profession and many characteristics of the customer, including health status and number of insurance claims, as well as the years the customer ordered the insurance and the amount spent on the insurance. Data details:

Table 1. Data details

applicant_id	
years_of_insurance_with_us	
regular_checkup_lasy_year	
adventure_sports	
Occupation	Salried
	Student
	Business
visited_doctor_last_1_year	
cholesterol_level	125 to 150
	150 to 175
	175 to 200
	225 to 250
daily_avg_steps	
age	
heart_decs_history	
other_major_decs_history	
Gender	male
	female
avg_glucose_level	
bmi	
smoking_status	never smoked
	Unknown
	formerly smoked
	smokes
Year_last_admitted	
weight	
covered_by_any_other_company	N
	Y
Alcohol	Rare
	No
	Daily
exercise	No
	Moderate
	Extreme
weight_change_in_last_one_year	
fat_percentage	
insurance_cost	

2.2.2. Data Preprocessing

Preprocessing data is a critical step that encompasses all the necessary actions taken prior to the modeling phase. It involves cleaning, transforming, and preparing raw data to make it suitable for training a machine learning model. First, let's understand the basic situation of the data:

```
In [11]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   applicant_id                    25000 non-null  int64
 1   years_of_insurance_with_us      25000 non-null  int64
 2   regular_checkup_lasy_year       25000 non-null  int64
 3   adventure_sports                25000 non-null  int64
 4   Occupation                      25000 non-null  object
 5   visited_doctor_last_1_year      25000 non-null  int64
 6   cholesterol_level               25000 non-null  object
 7   daily_avg_steps                 25000 non-null  int64
 8   age                             25000 non-null  int64
 9   heart_decs_history              25000 non-null  int64
 10  other_major_decs_history        25000 non-null  int64
 11  Gender                          25000 non-null  object
 12  avg_glucose_level               25000 non-null  int64
 13  bmi                             24010 non-null  float64
 14  smoking_status                  25000 non-null  object
 15  Year_last_admitted              13119 non-null  float64
 16  Location                        25000 non-null  object
 17  weight                          25000 non-null  int64
 18  covered_by_any_other_company    25000 non-null  object
 19  Alcohol                         25000 non-null  object
 20  exercise                        25000 non-null  object
 21  weight_change_in_last_one_year  25000 non-null  int64
 22  fat_percentage                  25000 non-null  int64
 23  insurance_cost                  25000 non-null  int64
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

Figure 1. Data's basic situation

First, delete the duplicate rows in the table. Duplicate rows are rows with exactly the same data. This kind of data has a bad impact on subsequent analysis, so it needs to be removed. After observation and processing, there are no duplicate rows in this data set.

Secondly, after observing the table above, it was found that there were missing values in bmi (24010numbers), And by observation, we can also find that there are many blank columns in the last admitted year. These blank columns in the original data represent no admitted. Here, the blank data is uniformly filled with the number 0 to represent no accident record. The data after this step of cleaning is as follow:

```
<class 'pandas.core.frame.DataFrame'>
Index: 24010 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   applicant_id                    24010 non-null  int64
 1   years_of_insurance_with_us      24010 non-null  int64
 2   regular_checkup_lasy_year       24010 non-null  int64
 3   adventure_sports                24010 non-null  int64
 4   Occupation                      24010 non-null  object
 5   visited_doctor_last_1_year      24010 non-null  int64
 6   cholesterol_level               24010 non-null  object
 7   daily_avg_steps                 24010 non-null  int64
 8   age                             24010 non-null  int64
 9   heart_decs_history              24010 non-null  int64
 10  other_major_decs_history        24010 non-null  int64
 11  Gender                          24010 non-null  object
 12  avg_glucose_level               24010 non-null  int64
 13  bmi                             24010 non-null  float64
 14  smoking_status                  24010 non-null  object
 15  Year_last_admitted              24010 non-null  float64
 16  Location                        24010 non-null  object
 17  weight                          24010 non-null  int64
 18  covered_by_any_other_company    24010 non-null  object
 19  Alcohol                         24010 non-null  object
 20  exercise                        24010 non-null  object
 21  weight_change_in_last_one_year  24010 non-null  int64
 22  fat_percentage                  24010 non-null  int64
 23  insurance_cost                  24010 non-null  int64
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

Figure 2. data's situation after the 2nd step of cleaning

Because the location is too messy and complicated, and this paper does not involve questions about the location, the location column data is deleted. Finally, in order to facilitate subsequent data processing, I converted the data as follows:

(1) Convert string type to floating point type, including occupations (Salried' as '0.0', 'Student' as '1.0', 'Business' as '2.0'), cholesterol level ('125 to 150' as '0.0', '150 to 175' as '1.0', '175 to 200' as '2.0', '200 to 225' as '3.0', '225 to 250' as '4.0'), Gender ('Male' as '0.0', 'Female': as 1.0'), smoking status ('Unknown' as '0.0', 'never smoked' as '1.0', 'formerlysmoked' as '2.0', 'smokes' as '3.0'), covered_by_any_other_company ('N'as'0.0', 'Y' as '1.0'), Alcohol ('No' as '0.0', 'Rare' as '1.0', 'Daily' as '2.0'), exercise ('No' as '0.0', 'Moderate' as '1.0', 'Extreme' as 2.0).

(2) To facilitate subsequent modeling, all data are converted to floating point types.

After this step is completed, the result is as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24010 entries, 0 to 24009
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   applicant_id                    24010 non-null  float64
 1   years_of_insurance_with_us      24010 non-null  float64
 2   regular_checkup_lasy_year       24010 non-null  float64
 3   adventure_sports                24010 non-null  float64
 4   Occupation                      24010 non-null  float64
 5   visited_doctor_last_1_year      24010 non-null  float64
 6   cholesterol_level               24010 non-null  float64
 7   daily_avg_steps                 24010 non-null  float64
 8   age                             24010 non-null  float64
 9   heart_decs_history              24010 non-null  float64
 10  other_major_decs_history        24010 non-null  float64
 11  Gender                          24010 non-null  float64
 12  avg_glucose_level               24010 non-null  float64
 13  bmi                             24010 non-null  float64
 14  smoking_status                  24010 non-null  float64
 15  Year_last_admitted              24010 non-null  float64
 16  weight                          24010 non-null  float64
 17  covered_by_any_other_company    24010 non-null  float64
 18  Alcohol                         24010 non-null  float64
 19  exercise                        24010 non-null  float64
 20  weight_change_in_last_one_year  24010 non-null  float64
 21  fat_percentage                  24010 non-null  float64
 22  insurance_cost                  24010 non-null  float64
dtypes: float64(23)
memory usage: 4.2 MB
```

Figure 3. Data's situation after the cleaning

At last all data are float type and got 24010 pairs of data. In order to avoid the impact of certain features on subsequent processing, the data is standardized and saved. The standardized results are as follows (Use z-score normalization method).

```
In [5]: scaled_df
```

Out[5]:

	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_
0	-1.731327	-0.416131	0.188493	3.347597	-1.631208	-0.967111	-1.003903	-0.3
1	-1.731189	-1.566540	-0.644512	-0.298722	-0.282819	0.782660	-0.211799	1.1
2	-1.731050	-1.183071	-0.644512	-0.298722	1.065570	0.782660	1.372407	-0.6
3	-1.730912	1.117748	2.687510	-0.298722	1.065570	-0.967111	0.580304	0.9
4	-1.730773	-0.416131	0.188493	-0.298722	-0.282819	-0.967111	-0.211799	-0.2
...	...	...	...	...	...	...	...	
24005	1.731003	1.117748	-0.644512	-0.298722	1.065570	-0.092226	-0.211799	0.2
24006	1.731142	-0.416131	-0.644512	-0.298722	-1.631208	0.782660	2.164510	0.3
24007	1.731281	0.734278	-0.644512	-0.298722	1.065570	0.782660	1.372407	-0.4
24008	1.731558	-1.183071	-0.644512	-0.298722	-1.631208	-0.967111	2.164510	5.2
24009	1.731696	1.501218	1.021499	-0.298722	1.065570	0.782660	-0.211799	0.6

24010 rows × 23 columns

Figure 4. standardized data

2.2.3. Feature selection and model selection

(1) Filter relevant data by calculating the Pearson correlation coefficient (PCC) and perform unsupervised learning.

	fat_percentage	insurance_cost
applicant_id	0.010627	-0.001386
years_of_insurance_with_us	-0.003078	0.001385
regular_checkup_lasy_year	0.002975	-0.175484
adventure_sports	0.003735	0.074257
Occupation	0.257733	0.009513
visited_doctor_last_1_year	-0.044901	0.008267
cholesterol_level	0.092064	-0.003025
daily_avg_steps	0.046580	-0.005218
age	-0.011447	0.004289
heart_decs_history	0.000339	0.001423
other_major_decs_history	0.004615	-0.001334
Gender	0.003317	0.002297
avg_glucose_level	0.002053	-0.003942
bmi	-0.003176	-0.008231
smoking_status	-0.009853	-0.008583
Year_last_admitted	-0.002020	0.161588
weight	-0.006757	0.970460
covered_by_any_other_company	0.005755	0.101953
Alcohol	-0.017300	-0.006547
exercise	0.012147	0.002071
weight_change_in_last_one_year	0.012818	-0.341177
fat_percentage	1.000000	-0.008072
insurance_cost	-0.008072	1.000000

Figure 5. Data's Pearson correlation coefficient (PCC)

(2) Use visualization tools to draw graphs and observe the connections between data

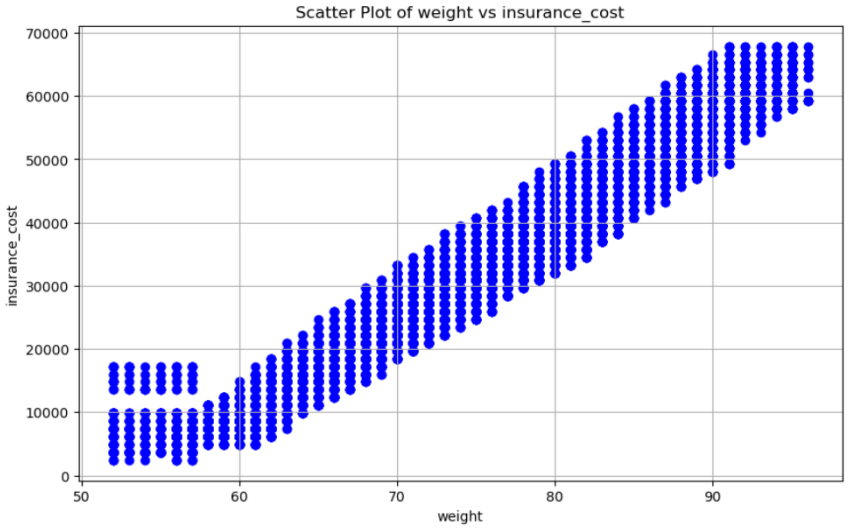

Figure 6. Scatter plot of weight and insurance cost

(3) Based on the customer's own characteristics and ideas, and using the marketing logic in the real world, we can determine the possible correlation between data pairs.

After a variety of quantitative and qualitative screening methods, the following data were finally selected for final modeling analysis:

a. Weight and insurance cost: In the Pearson correlation coefficient, the correlation between weight and insurance costs is very high, and it is also reflected in the scatter plot. In the logic of daily life, it is also very likely that people with heavier weight will buy a lot of commercial medical insurance to prevent their health problems, because obesity often causes more diseases. Therefore, the first group of analyses uses weight and insurance costs for unsupervised learning.

b. Smoke and alcohol: As we all know, drinking and smoking can have many adverse effects on our health. After suffering from chronic or acute diseases, many people may purchase commercial health insurance or commercial medical insurance to protect their rights and convenience of medical

treatment. Therefore, based on this possible situation in reality, I extracted smoking and alcohol as a set of feature values.

2.2.4. Modeling

Unsupervised learning, the machine simply receives inputs x1,x2,..., but obtains neither supervised target outputs, nor rewards from its environment. It may seem somewhat mysterious to imagine what the machine could possibly learn given that it doesn't get any feedback from its environment. However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise. Two very simple classic examples of unsupervised learning are clustering and dimensionality reduction. In this article, we mainly discuss the first type of problem: clustering

(1) Since unsupervised learning does not set a goal in advance like supervised learning, and has the distinction between training set and test set, unsupervised learning pays more attention to the nature and internal rules of data. Therefore, when modeling, I use weight and insurance cost as two variables and perform K-mean algorithm in unsupervised learning (the first item above), n_init is set to 20 (In order to obtain better clustering results, the algorithm usually performs multiple random initializations, each time using a different random seed (random state) to generate the initial centroids. The n_init parameter specifies the number of these random initializations.); The number of clusters k is set to 3; The maximum number of iterations is 300. The results are as follows:

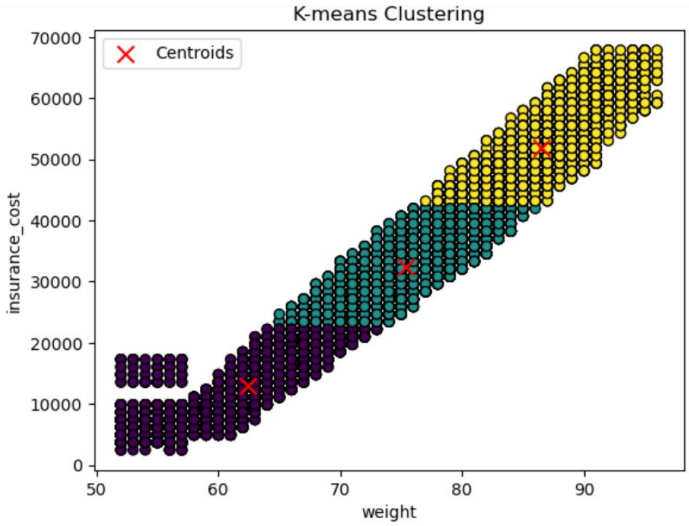

Figure 7. K-means clustering of weight vs insurance cost

(2) For the second set of data, I plan to use the association rule algorithm. Here are some algorithmic models that I might use for analysis:

Apriori algorithm: a classic frequent itemset mining algorithm, suitable for processing large-scale data sets.

FP-Growth algorithm: uses the Frequent Pattern Growth method to mine frequent itemsets, which is usually more efficient than Apriori.

Eclat algorithm: a vertical data format based on searching for candidate itemsets, suitable for processing dense data sets.

After careful consideration, I chose the Apriori analysis method. The following is the training code and visualization results:

```
In [2]: data = pd.read_csv('data_2.csv')
        data = pd.DataFrame(data)

In [3]: # Convert Alcohol and smoking_status to Boolean values
        data['Alcohol'] = data['Alcohol'].apply(lambda x: True if x >= 1 else False)
        data['smoking_status'] = data['smoking_status'].apply(lambda x: False if x == 1 else True)

In [4]: # Convert to string type
        data['Alcohol'] = data['Alcohol'].astype(str)
        data['smoking_status'] = data['smoking_status'].astype(str)

In [5]: # Transaction data converted into a list
        transactions = []
        for idx, row in data.iterrows():
            transaction = [row['Alcohol'], row['smoking_status']]
            transactions.append(transaction)

        # One-hot encoding using TransactionEncoder
        te = TransactionEncoder()
        te_ary = te.fit_transform(transactions)
        df = pd.DataFrame(te_ary, columns=te.columns_)

In [6]: # Use Apriori algorithm to find frequent itemsets
        frequent_itemsets = apriori(df, min_support=0.2, use_colnames=True)

In [7]: # Visualizing Frequent Itemsets
        plt.figure(figsize=(10, 6))
        plt.barh(range(len(frequent_itemsets)), frequent_itemsets['support'], align='center', alpha=0.8)
        plt.yticks(range(len(frequent_itemsets)), frequent_itemsets['itemsets'].apply(lambda x: ', '.join(x)))
        plt.xlabel('Support')
        plt.ylabel('Frequent Itemsets')
        plt.title('Frequent Itemsets Found by Apriori Algorithm')
        plt.gca().invert_yaxis()
        plt.show()

        # Generate association rules
        rules = association_rules(frequent_itemsets, metric='lift', min_threshold=0.5)

        # Output association rules
        print("Association Rules:")
        print(rules)
```

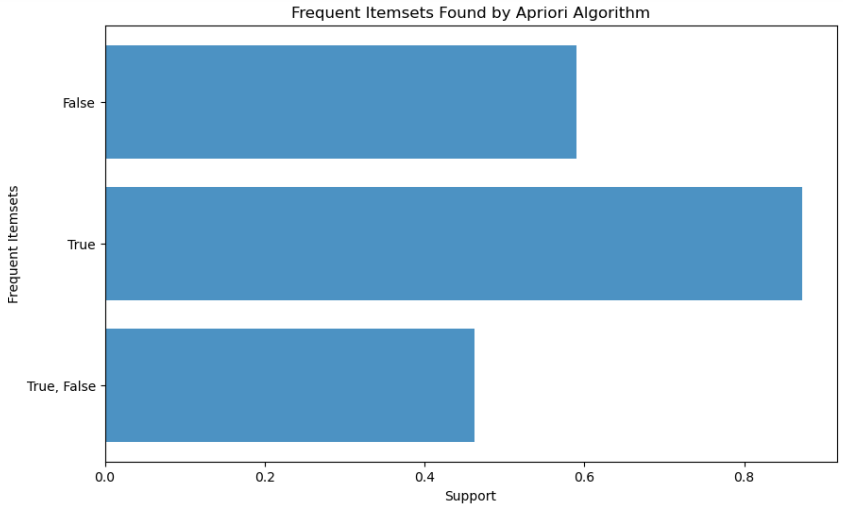

Figure 8. programming and results of Apriori Algorithms

2.2.5. Performance metrics

For the first model(weight and insurance cost), there is a strong linear correlation overall. The data results at this time are basically equivalent to linear regression, and the data is divided into three categories. However, for the sake of rigor, I still selected many performance indicators (including distortion degree, adjusted mutual information, silhouette coefficient, and adjusted Rand index) to test the model.

Introduction to Silhouette Analysis

Silhouette analysis is a method used to evaluate the quality of clusters created by a clustering algorithm, such as K-means. It measures how similar an object is to its own cluster compared to other clusters. The silhouette score for each sample is a value between -1 and 1:

(1) A score close to 1 indicates that the sample is well-matched to its own cluster and poorly matched to neighboring clusters.

(2) A score close to 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters.

(3) A negative score indicates that the sample might have been assigned to the wrong cluster.

The average silhouette score of all samples provides an overall measure of how well the clustering has been performed. Higher average scores indicate better-defined clusters.

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

plt.figure(figsize=(8, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=y_kmeans, s=50, cmap='viridis', marker='o', edgecolor='k')
centroids = kmeans.cluster_centers_
centroids_pca = pca.transform(centroids)
plt.scatter(centroids_pca[:, 0], centroids_pca[:, 1], c='red', s=200, alpha=0.75, marker='x', label='Centroids')
plt.title('K-means Clustering with PCA')
plt.xlabel('PCA Feature 1')
plt.ylabel('PCA Feature 2')
plt.legend()
plt.grid(True)
plt.show()
```

```
Silhouette Coefficient: 0.5272211743721736
```

Silhouette Coefficient: 0.5272211743721736

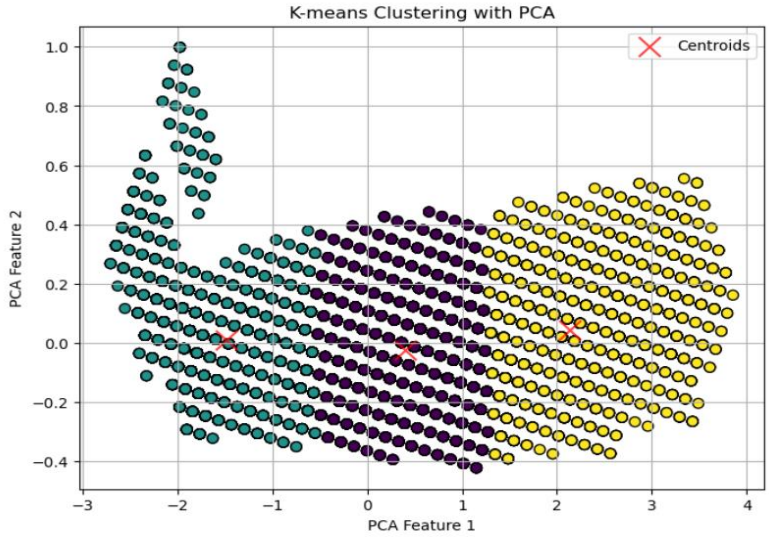

Figure 9. programming and results of silhouette score and K-mean clustering with PCA

The silhouette coefficient is 0.53, indicating that the sample has a certain degree of match with itself and its neighbors, and such a model is acceptable.

For the second one,The confidence of the model is shown in the figure below:

```
print('Association rules and confidence:')
print(rules[['antecedents', 'consequents', 'confidence']])
```

```
Association rules and confidence:
  antecedents consequents  confidence
0      (True)     (False)    0.530684
1     (False)      (True)    0.783473
```

Figure 10. confidence of the model

Confidence in association rule mining denotes the measure of how likely it is for the consequent (outcome) to occur when the antecedent (condition) is satisfied. It quantifies the frequency with which the items in the consequent appear in transactions that include the items in the antecedent. Higher confidence values indicate a stronger statistical relationship between the antecedent and consequent items. Confidence ranges from 0 to 1, where a value closer to 1 signifies a higher probability that the consequent will occur given the antecedent, thus suggesting a more reliable association rule.

Based on this, this model is completely acceptable.

3. RESULTS AND DISCUSSION

In this study, I collected a total of 25,000 real insurance sales data. After data cleaning and other steps, I finally left 24,010 samples, which contained many characteristics of customers who bought insurance (insurance costs, weight, smoking and drinking, etc.). I used a variety of methods and analyses to select more significant features and performed unsupervised learning (Kmean and association rule learning) on them. Finally, after multiple modifications and evaluations of the model, I obtained a relatively satisfactory result. The following will be explained in detail:

For the first model, All indexes are relatively normal and belong to qualified models.

For the second one, The confidence value measures the probability of the subsequent item appearing under the condition that the preceding item appears. For example, the confidence value in the first rule is 0.530684, and the confidence value in the second rule is 0.783473. Generally speaking, the higher the confidence value, the more reliable the rule. It is necessary to understand whether the meaning of the rule is reasonable and practical. For example, the first rule points out that under the condition that the preceding item is True, the probability of the subsequent item being False is 53.07%. Whether this information is meaningful in the actual business scenario needs to be determined based on the specific situation. Although no support information is provided, support is also an indicator for evaluating the importance of association rules. Rules with high support are usually more universal and influential. Based on the above points, it can be preliminarily judged that these association rules may be qualified in some cases, but the final qualification evaluation needs to be further confirmed and adjusted in combination with specific business needs and actual application scenarios.

4. CONCLUSION

After discussing the model, let us move on to the next section and discuss the results from the perspective of the insurance company's specific sales business, including the policies and suggestions to be implemented.

In insurance sales, strategizing based on customer weight can significantly influence consumer behavior and policy uptake. Understanding how weight affects insurance premiums and coverage options allows agents to tailor recommendations that resonate with clients' health concerns and financial goals. By highlighting the correlation between healthy weight management and long-term insurance benefits, agents can emphasize policies that offer incentives for maintaining healthy lifestyles. This proactive approach not only fosters trust but also positions insurance products as proactive investments in overall well-being. Thus, integrating weight considerations into sales strategies not only enhances customer engagement but also promotes a holistic approach to insurance planning.

In the first model, it can be clearly seen that customers are divided into three groups. We can observe that the heavier the person, the greater the cost of insurance. Therefore, in insurance companies, managers and salespeople can develop some more targeted insurance, such as medical insurance that protects many diseases that overweight people are prone to, and then conduct very targeted sales for people with heavier weight (the data in this model is 80kg, and the specific situation needs to be determined according to the local population). There will definitely be a significant increase in sales value. Secondly, because the data is divided into three obvious categories, for people with lighter weight, in the final presentation of the results, their insurance costs may be lower, but the salesperson can change his way of thinking to think about the problem, which leads to the second model.

Since the data collected is a sales data, in association rule learning, unlike normal supermarket shopping, where you buy several items at the same time, in this model, I regard buying insurance as going to the supermarket, so all the data in the data set belongs to this category. Secondly, I selected drinking and smoking as two important indicators that significantly affect whether people buy

commercial medical insurance, and modeled them. The final result shows that the support for the two items appearing true at the same time reached an astonishing 0.8, and the true and false cases also accounted for half.

There is no doubt that people who have the habit of smoking and drinking also buy insurance, and half of people who have one of the two habits also buy insurance. For marketers, it is very necessary to design and promote targeted insurance types for people who have both or one of the above two habits, because they are really likely to spend money on insurance.

REFERENCES

[1] Nyaga, P. K., & Muema, M. W. (2017). AN ANALYSIS OF THE EFFECT OF PRICING STRATEGIES ON PROFITABILITY OF INSURANCE FIRMS IN KENYA. International Journal of Finance and Accounting, 2(3), 44–65. https://doi.org/10.47604/ijfa.319.

[2] Gao, J.X., Zhou, Z.R., Ai, J.S., Xia, B.X. and Coggeshall, S. (2019) Predicting Credit Card Transaction Fraud Using Machine Learning Algorithms. Journal of Intelligent Learning Systems and Applications, 11, 33-63.

[3] Moon, H., Pu, Y. and Ceglia, C. (2019) A Predictive Model ing for Detecting Fraudulent Automobile Insurance Claims. Theoretical Economics Letters, 9, 1886-1900. https://doi.org/10.4236/tel.2019.96120

[4] Mare, C.; Manaţe, D.; Mureşan, G.-M.; Dragoş, S.L.; Dragoş, C.M.; Purcel, A.-A. Machine Learning Models for Predicting Romanian Farmers' Purchase of Crop Insurance. Mathematics 2022, 10, 3625. https://doi.org/10.3390/math10193625

[5] EL KOUFI, N. O. U. H. A. I. L. A., and ABDESSAMAD BELANGOUR. "RESEARCH INTELLIGENT PRECISION MARKETING OF INSURANCE BASED ON EXPLAINABLE MACHINE LEARNING: A CASE STUDY OF AN INSURANCE COMPANY." Journal of Theoretical and Applied Information Technology 102.6 (2024).

[6] Goundar, S., Prakash, S., Sadal, P., & Bhardwaj, A. (2020). Health insurance claim prediction using artificial neural networks. International Journal of System Dynamics Applications, 9(3), 40-57. https://doi.org/10.4018/ijsda.2020070103

[7] Chandrasekaran, S. and Kumar, A. (2019) A Clustering Ap proach for Customer Billing Prediction in Mall: A Machine Learning Mechanism. Journal of Computer and Communications, 7, 55-66. https://doi.org/10.4236/jcc.2019.73006

[8] Kaiser Family Foundation. (2023, June 15). Americans' challenges with health care costs. KFF. https://www.kff.org/report/americans-challenges-health-care-costs/

[9] Peterson-KFF Health System Tracker. (2022, December 21). The state of the U.S. health system in 2022 and the outlook for 2023. Health System Tracker. https://www.healthsystemtracker.org/brief/the-state-of-the-u-s-health-system-in-2022-and-the-outlook-for-2023/

[10] Kaiser Family Foundation. "Americans' Challenges with Health Care Costs." KFF, 15 June 2023. https://www.kff.org/report/americans-challenges-health-care-costs/

[11] Peterson-KFF Health System Tracker. "The State of the U.S. Health System in 2022 and the Outlook for 2023." Health System Tracker, 21 December 2022. https://www.healthsystemtracker.org/brief/the-state-of-the-u-s-health-system-in-2022-and-the-outlook-for-2023/

[12] García, S., Ramírez-Gallego, S., Luengo, J. et al. Big data preprocessing: methods and prospects. Big Data Anal 1, 9 (2016). https://doi.org/10.1186/s41044-016-0014-0

[13] Ghahramani, Z. (2004). Unsupervised Learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds) Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science(), vol 3176. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_5

[14] Rousseeuw, P. J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics*, 20: 53-65. DOI: 10.1016/0377-0427(87)90125-7.

[15] Hahsler, M., Chelluboina, S., & Hornik, K. (2007). "The arules R-package ecosystem: Analyzing interesting patterns from large transaction datasets." *Journal of Machine Learning Research*, 12: 2021-2025. Link