

Multiple algorithms helping insurance companies find customers to sell car insurance

Team: 斗帝组

Team members: 毛瑞, 韩光茜, 王一文, 郝颖

I. Introduction

1. Problem

Generally speaking, for a company, the more products it sells, the higher its profit will be. In order to obtain more profit, a certain health insurance company in the United States customized various health insurances for its clients. An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee. In order to expand the company's business, the company introduced a new type of Vehicle Insurance and promoted it to its customers. Just like medical insurance, the vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called 'sum assured') to the customer.

The company recorded the customers who bought the auto insurance from this company and their information for several years. We will use these customer information for data mining, analyze customer information to establish a predictive model to determine whether a customer will be willing to buy auto insurance, and then visually view customer data through a visual model, and use cluster analysis and association analysis to determine the purchase of the insurance customers are classified to assign appropriate salespersons. The above can

greatly reduce the company's labor costs and sales success rate, thereby increasing the company's profitability.

2. Data

The data is provided on “Kaggle”, which is a web site dedicated to providing data information. The company given the personal information of 508,147 customers who purchased Vehicle Insurance, which included eleven items(Figure 1) such as the age of the customer, the age of the car, and whether the car was damaged and so on. We all use them as the attributes for searching. The customer's purchase intention is represented by the Numbers “0” and “1”, “1” means willing to buy automobile insurance, “0” means unwilling.

Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Driving_License	0 : Customer does not have DL, 1 : Customer already has DL
Region_Code	Unique code for the region of the customer
Previously_Insured	1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
Vehicle_Age	Age of the Vehicle
Vehicle_Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
Annual_Premium	The amount customer needs to pay as premium in the year
PolicySalesChannel	Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
Vintage	Number of Days, Customer has been associated with the company
Response	1 : Customer is interested, 0 : Customer is not interested

(Figure 1)

The identity information of the customers is made into two data sets, and the two data sets record the information of different customers. One data set is the training set (training.csv) with 381110 customers' information, and the other data set is the test set (test.csv) with 127037 customers' information.

II. Experiment

1. Goal

We used all the data for data visualization to analyze the distribution and characteristics of the data, and then used the data from the training set to build a model to predict the intention of insurance company customers to buy auto insurance, and found the group of customers willing to buy auto insurance through clustering and correlation analysis. In this way, the insurance company can predict the purchase intention of customers and carry out appropriate marketing, and arrange special salesmen to receive different customers.

2. Requirement

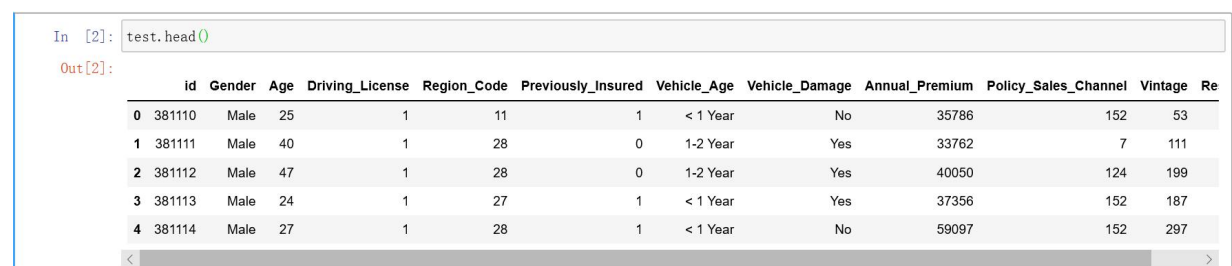
sklearn numpy pandas time texttable
Seaborn scikit-learn matplotlib plotly

3. Algorithm

A. Visualization

Visualization can transform invisible data phenomena into visible graphic symbols. It can establish connections and associations between intricate and seemingly inexplicable data, discover laws and characteristics, and gain insights and values of more commercial value.

(1) Arranges the gibberish of ID Numbers into an ordered sequence. For example, view first 5 data rows.(Figure 2)



```
In [2]: test.head()
```

Out[2]:

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Re
0	381110	Male	25	1	11	1	< 1 Year	No	35786	152	53	
1	381111	Male	40	1	28	0	1-2 Year	Yes	33762	7	111	
2	381112	Male	47	1	28	0	1-2 Year	Yes	40050	124	199	
3	381113	Male	24	1	27	1	< 1 Year	Yes	37356	152	187	
4	381114	Male	27	1	28	1	< 1 Year	No	59097	152	297	

In [4]:

train.head()

Out[4]:

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Respon
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0	26.0	27	
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0	152.0	203	
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0	152.0	39	

(Figure 2)

(2) See how the data is distributed and the maximums, the minimums, the mean, ...(Figure 3)

In [5]:

train.describe()

Out[5]:

	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
count	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000
mean	190555.000000	38.822584	0.997869	26.388807	0.458210	30564.389581	112.034295	154.347397	0.122563
std	110016.836208	15.511611	0.046110	13.229888	0.498251	17213.155057	54.203995	83.671304	0.327936
min	1.000000	20.000000	0.000000	0.000000	0.000000	2630.000000	1.000000	10.000000	0.000000
25%	95278.000000	25.000000	1.000000	15.000000	0.000000	24405.000000	29.000000	82.000000	0.000000
50%	190555.000000	36.000000	1.000000	28.000000	0.000000	31669.000000	133.000000	154.000000	0.000000
75%	285832.000000	49.000000	1.000000	35.000000	1.000000	39400.000000	152.000000	227.000000	0.000000
max	381109.000000	85.000000	1.000000	52.000000	1.000000	540165.000000	163.000000	299.000000	1.000000

(Figure 3)

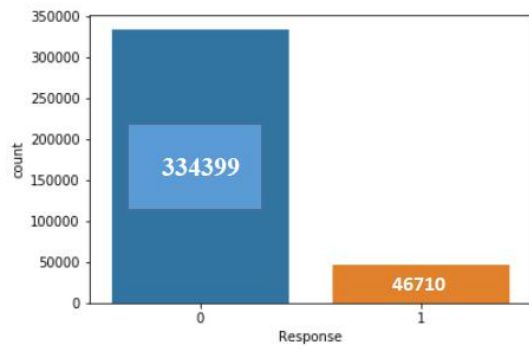
(3) See what type of data the information is.(Figure 4)

In [6]:	train.info()
	<pre> <class 'pandas.core.frame.DataFrame'> RangeIndex: 381109 entries, 0 to 381108 Data columns (total 12 columns): # Column Non-Null Count Dtype --- --- - 0 id 381109 non-null int64 1 Gender 381109 non-null object 2 Age 381109 non-null int64 3 Driving_License 381109 non-null int64 4 Region_Code 381109 non-null float64 5 Previously_Insured 381109 non-null int64 6 Vehicle_Age 381109 non-null object 7 Vehicle_Damage 381109 non-null object 8 Annual_Premium 381109 non-null float64 9 Policy_Sales_Channel 381109 non-null float64 10 Vintage 381109 non-null int64 11 Response 381109 non-null int64 dtypes: float64(3), int64(6), object(3) memory usage: 34.9+ MB </pre>

(Figure 4)

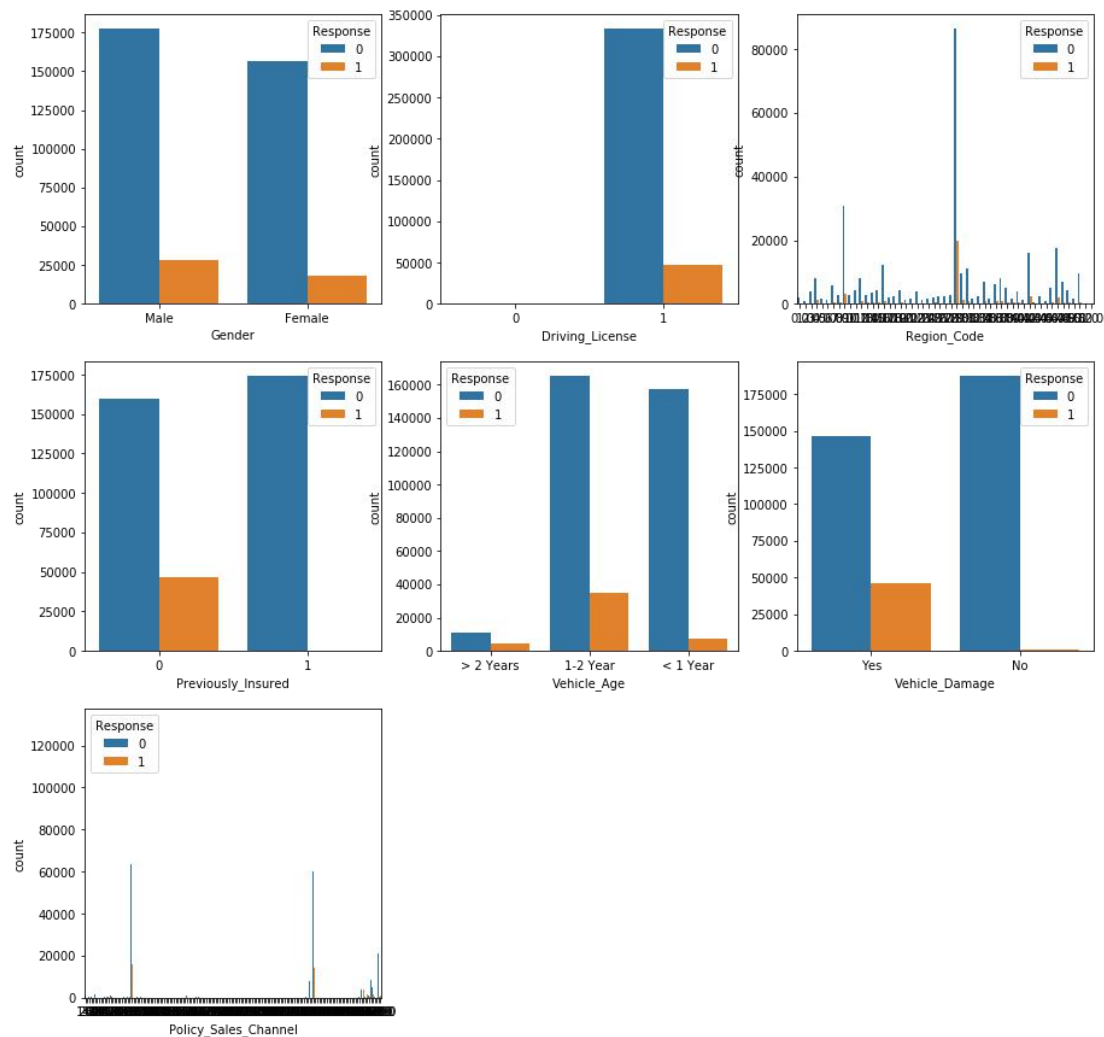
(4) Count of response

Code: sns.countplot(train.Response)



(Figure 5. “1” means customer willing to buy automobile insurance, “0” means unwilling.)

(5) Categorical features



(Figure 6. The orange bars represent people who buy car insurance, and the blue bars represent people who don't.)

From figure 6, many distinctions of the customers between who bought vehicle insurance and who didn't:

Gender: Male customers are more in number than female customers. Although male candidates have responded positively to the car insurance proposal more frequently than the female customers. It is proportional to the total number. Therefore, gender doesn't tell us a lot about the response.

Driving License: All the customers have a driving licence. Therefore, driving license is not a good predictor of the customers response. We might not even use Driving License as a feature in our prediction.

Region Code: Although Region code is a categorical feature. The number of categories are a lot, so it will be analyzed as a discrete numeric feature later.

Previously Insured: Majority of the customers are not previously insured(Vehicle Insurance). Almost all of the customers who have responded positively were not insured previously. This feature can be a great predictor of the customers response.

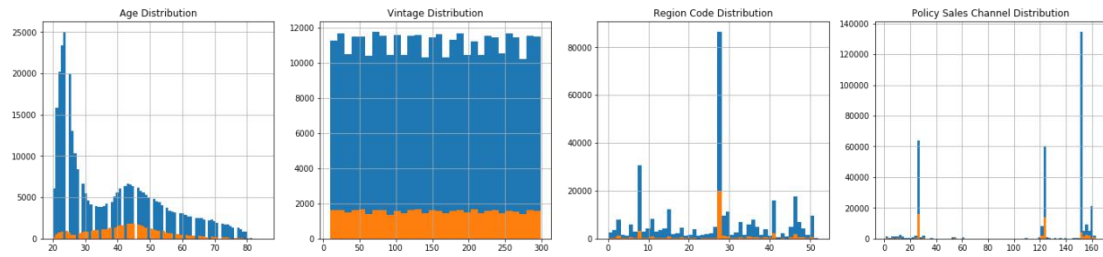
Vehicle Age: Majority of the customers have a vehicle age of ≤ 2 years. Customers with Vehicle of age 1-2 years are more likely to get an insurance. This feature will be useful in our predictions.

Vehicle Damage: The distribution of customers who have damaged their vehicles in the past is almost identical. Customers who have damaged their vehicles in the past are more likely to buy a vehicle insurance.

Policy_Sales_Channel: Just like Region Code number of categories are a lot in Policy Sales Channel, so it will be analyzed as a discrete numeric feature later.

(6) Numeric Features

Discrete numeric: Using buckets in histogram for better visualization.(Figure 7)



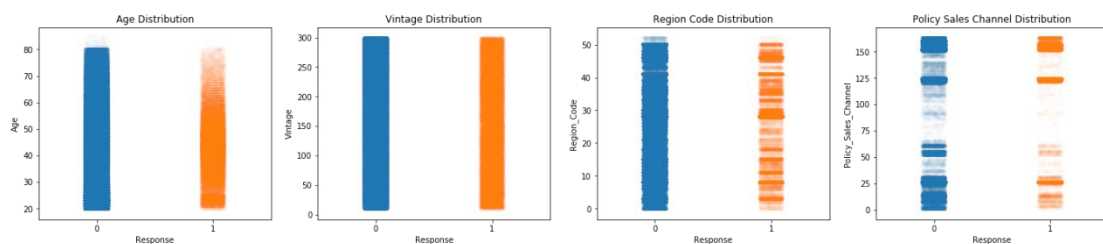
(Figure 7. The orange part represents people who buy car insurance, and the blue part represents people who don't.)

Figure 8 is a table with analysis of the visualization of the discrete numeric features.

Feature	Distribution	Description	Take Away
Age	Vaguely similar to Log-Normal	Majority of the customers are either young or close to retirement	Although young customers are more in number. Early retired age customers are more likely to purchase a car insurance
Vintage	Very close to Uniform	The customers are distributed uniformly	The number of days of association with the company has almost no effect on the likelihood of car insurance.
Region Code	Not a standard distribution	Regions 8, 28, 41, 46 have most of the customers	The likelihood of purchasing insurance is directly correlated to the number of customers in the regions.
Policy Sales Channel	Not a standard distribution	Spikes in channels of sales channels 26,124, 152-160 are most effective	Channels 26, 124 are relatively more successful in selling car insurance.

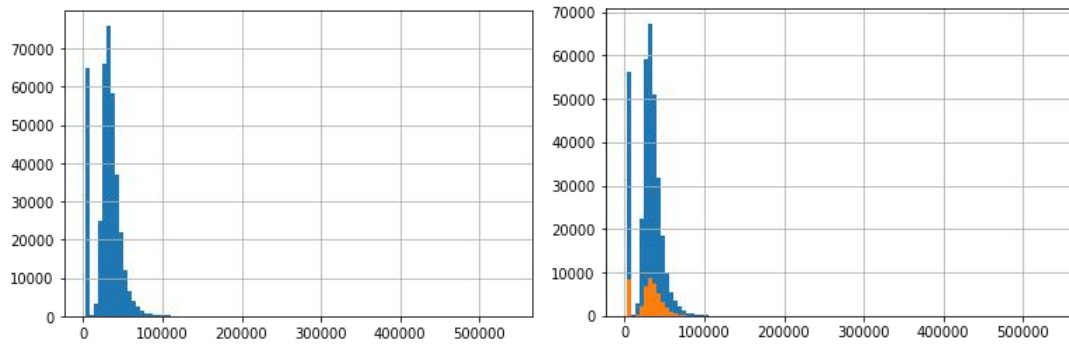
(Figure 8)

Outlier detection for figure 7 (Figure 9) . We can see there is no visible outlier in the discrete numeric attributes.



(Figure 9)

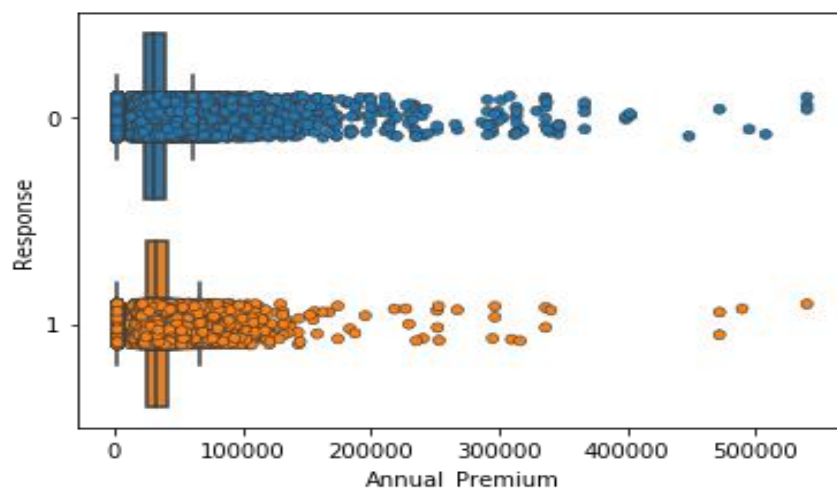
Continuous numeric: The annual premium also contributes to the customers' idea.(Figure 10)



(Figure 10. The orange part represents people who buy car insurance, and the blue part represents people who don't.)

Take Away: The likely hood of purchasing insurance is directly correlated to the number of customers in the particular annual premium group. Most likely to have outliers.

Outlier detection of figure 10 (Figure 11). Annual Premium has a lot of outliers. These outliers will affect the values our model will learn and will lead to skewed predictions. Let's keep this in mind, we might drop the outliers (customers with annual premium more than 200000) if needed.



(Figure 11)

B. Classification

Classification is to find out the common characteristics of a group of data objects in the database and divide them into different classes according to the classification pattern. Its

purpose is to map the data items in the database to a given category through the classification model.

(1) Establish the classifier and model

To help this insurance companies to find the right customers to buy vehicle insurance, we build models for four classifiers including Decision Tree, KNN, MLP. Then, we build ensembles using above models. Finally, we can make prediction with model that has best accuracy. When building models, we set different parameters for some of classifiers to help better training according to given data sets.

a. **Decision Tree:** We let $\{\text{max_depth}=29, \text{min_samples_split}=350\}$. max_depth limits the depth of tree. min_samples_split limits samples used for training. Setting the two values can help to avoid overfitting.

b. **KNN:** We let $\{\text{n_neighbors}=30, \text{weights}='uniform'\}$. We choose "uniform" for weights so that all nearest neighbor samples have the same weight. As for the value of n_neighbors , we find that it is quite difficult to decide. Although the training accuracy is quite high, the training time is quite long. Thus, we use a tool GridSearchCV to help us find the best parameters. To use this tool, we just need to input the classifier, values of parameters except the best parameter need to find, and the possible values of the best parameter. Just like the picture as follows. Besides, the parameter scoring is criteria of model evaluation, the parameter "cv" is a cross-validation parameter, the parameter n_jobs are number of processes, and we make it 1 to be easy to adjust.

c. **MLP:** We let $\{\text{solver}='adam', \text{batch_size}=1000, \text{max_iter}=500, \text{beta_1}=0.85, \text{beta_2}=0.7\}$. The default solver "adam" is chosen since it performs well in terms of training time and validation scores for large data sets. Batch_size decides the size of the minibatch used for random optimizer. Max_iter stands for the maximum number of iterations. During several training processes, I find that making max_iter small can help decrease training time. Beta stands for the exponential decay rate of the moment vector.

d. **Boosting:** We let `{base_estimator=None, learning_rate=1.0, n_estimators=50, algorithm='SAMME.R', random_state=None}`. The learning rate indicates the convergence speed of the gradient. If it is too large, it will be easy to miss the optimal value. If it is too small, the convergence speed will be very slow and easy to underfit. `N_estimators` indicates base classifier's cycle times. If this value is too large, the model may be overfitting, and if the value is too small, the model may be underfitting. Algorithm indicates model's promotion guidelines, and SAMME.R use the proportion of prediction errors in the sample set to be the guideline.

e. **Bagging:** we let `{n_estimators=50, max_samples=150, max_features=10, bootstrap=True, n_jobs=-1}`. `N_estimators` indicates number of basic estimator in the set, and `max_samples` indicates the number of samples drawn to train each basic estimator. `Bootstrap=True` indicates that we will replace samples.

Here are the test of each classification model (Figure 12):

classifier	parameters	accuracy	training time
DT	<code>{'max_depth': 29, 'min_samples_split': 350}</code>	0.983	2.510
K-NN	<code>{'n_neighbors': 30, 'weights': 'uniform'}</code>	1.000	35.477
mlp	<code>{'solver': 'adam', 'batch_size': 1000, 'max_iter': 500, 'beta_1': 0.85, 'beta_2': 0.7}</code>	0.990	131.233
Adaboost	<code>{'base_estimator': 'None', 'learning_rate': 1.0, 'n_estimators': 50, 'algorithm': 'SAMME.R', 'random_state': 'None'}</code>	1	18.023
Bagging	<code>{'n_estimators': 50, 'max_samples': 150, 'max_features': 10, 'bootstrap': 'True', 'n_jobs': -1}</code>	0.997	1.874

(Figure 12)

(2) Comparison and discussion

By testing each classifier several times, we get the table shown in Figure 12, from which we can make a comparison between several classifiers. And the comparison is shown in figure 13 below.

Classifier	Advantage	Disadvantage
Decision Tree	Visualization, easy to understand; handle interaction between features easily	Easy to be overfitting
KNN	Can handle nonlinear classification; high accuracy; not sensitive to outliers	long calculation; misjudge with unbalanced sample classification
MLP	Can access information from large data sets and construct complicated models.	Require careful data processing; more training time
Adaboost	Easy to build; not easy to overfitting; no requirement to filter features; reduce bias	Sensitive to outliers
Bagging	Reduces the variance	

(Figure 13)

Figures 12 and 13 show us the performance of each classifier:

- a. For Decision Tree, the advantage is that it can visualize classification ways by the tree with leaf nodes. Also, it can handle the interaction between features easily, so that we don't have to worry about outliers or whether the data is linearly separable. However, the disadvantage is that if the depth of the tree is not limited, overfitting may occur since the depth of the tree will be so large that it remembers all labels of the training data. Therefore, we set `max_depth` or `min_samples_split` to prune.
- b. For KNN, the special one is that its training is sensitive to training data sets, especially the value of `n_neighbor`. When the neighbor value is too small, it uses a small neighborhood for prediction. If the neighbor happens to be a noisy point, it will lead to overfitting. As the neighbor value continues increasing, the training accuracy of the model first increases and then decreases, and the turning point is the neighbor value we are looking for. Its advantage is that it can be used for nonlinear classification, it has high accuracy, and it is not sensitive to

outliers. However, its disadvantages are that the amount of its calculation is too large, and there may be misjudgments with unbalanced sample classification.

c. For MLP, its advantage is that it can access information from large data sets and construct complicated models. However, it needs careful data pre-processing and parameters adjusting. Besides, it usually requires more training time. To adjusting parameters, the first thing is to create a network which is large enough to overfit to ensure that the network can learn the task. Then, shrink the network or increase the alpha to enhance the regularization, which can improve the generalization performance. After this, for those who chooses adam to be solver, set the batch size, the number of iterations and the beta value. In my MLP training process, it takes much more time to adjust parameters compared to other classifiers.

d. For Adaboost, its model is built based on the error rate of the previous model. It will pay more attention to the wrong samples, and reduce the attention to the correctly classified samples. After successive iterations, we can obtain a relatively good model. The advantage is that we can different ways to build sub-classifier using the framework Adaboost provides. The model is easy for us to build and we do not need to filter features. Also, it will not be easy to be overfitting. However, its disadvantage is that it is sensitive to outliers.

e. For Bagging, it averages the results of every model' s accuracy. Compared with a single decision tree, it reduces the variance of the basis estimator by introducing randomness in the process of building the model and has a smoother classification boundary. In our bagging training process, we find that increasing n_estimator value can improve training accuracy.

In general, the performance of these classifiers is very good, each classifier has its own advantages, they can accurately determine whether the customer is willing to buy car insurance according to the information.

In real life, we will recommend these different algorithms to the company at the same time, and let the company make the final decision.

C. Logistic regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. It gives the probabilistic values which lie between 0 and 1, and is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values ("0" or "1"). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Steps in Logistic Regression:

- Data Pre-processing step
- Fitting Logistic Regression to the Training set
- Predicting the test result
- Test accuracy of the result (Creation of Confusion matrix)
- Visualizing the test set result.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete data sets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

D. Clustering-K means

K-means is one method of cluster analysis that groups observations by minimizing Euclidean distances between them.

The algorithm works as follows:

1. First initialize k points, called means, randomly.
2. Categorize each item to its closest mean and update the mean's coordinates, which are the average of the items categorized in that mean so far.
3. Repeat the process for a given number of iterations and obtain the clusters.

Feature choice:

In this project, through the beforehand visualization process, features 'Gender', 'Driving License', 'Vintage' has no obvious relationship between 'Response', i.e., the interest to buy auto insurance, and features 'Previously Insured', 'Vehicle Age', 'Vehicle Damage' has 0-1 distribution or 0-1-2 distribution, which means we can also derive from the visualization process that what kind of feature value would result in a positive response, and vice versa. For instance, almost all of the customers who have responded positively were not insured previously, and customers who have damaged their vehicles in the past are more likely to buy a vehicle insurance and customers with Vehicle of age 1-2 years are more likely to get an insurance. In the end, we have 'Age', 'Annual Premium', 'Region Code', 'Policy Sales Channel' as our feature candidates.

Through testing, we find that feature 'Annual Premium' didn't give us a satisfactory result in that we were unable to combine it with any other feature to form a cluster, and we gone back to refer to the visualization diagrams, finding that the likelihood of purchasing insurance is directly correlated to the number of customers in the particular annual premium group. So we drop this feature as well due to its dominant effect.

K choice: The elbow method

Key indicator: SSE (Sum of the squared errors)

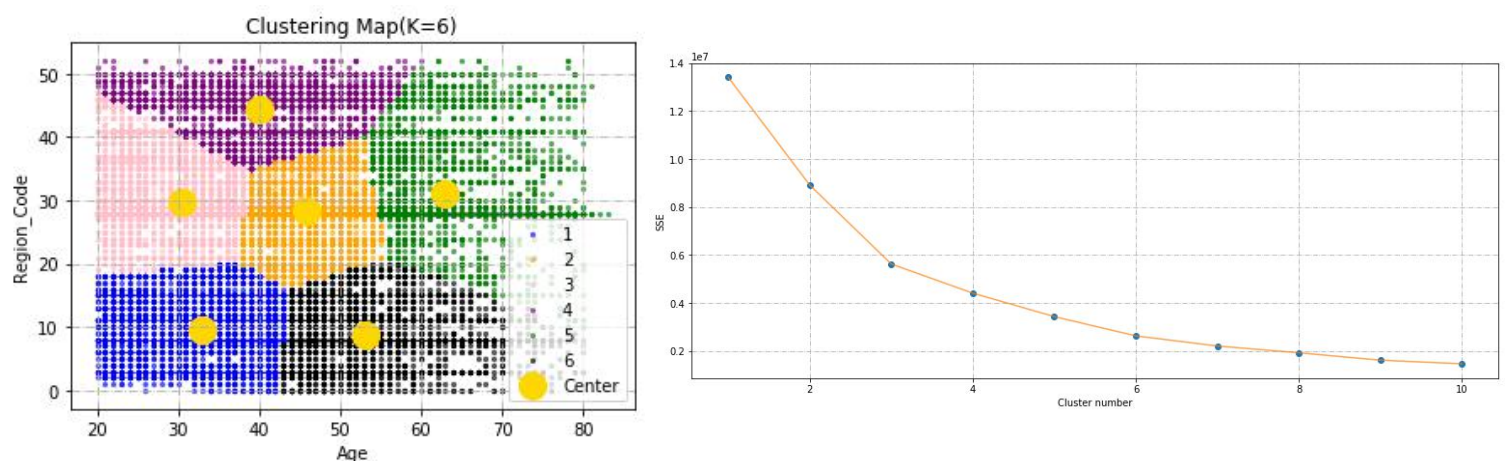
Main idea:

With the increase of the clustering number k , the sample division will become more refined, and the degree of aggregation of each cluster will gradually increase, so the error square sum SSE will naturally gradually decrease.

When k is less than the true cluster number, due to the increase of k , the degree of polymerization of each cluster will substantially increase, so the SSE will drop accordingly, and when k arrived at the real cluster number, adding k will result in smaller return of polymerization degree, so SSE drop will decrease and tend to flatten out with k value continues to increase, that is to say, SSE- k diagram is in the form of an elbow, and the corresponding k value of the elbow is the real cluster number.

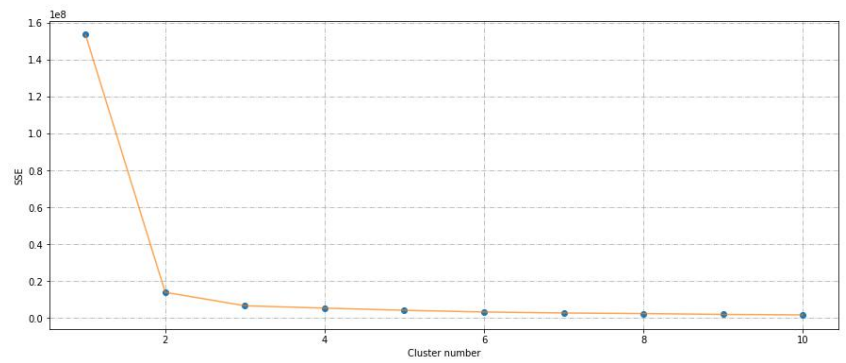
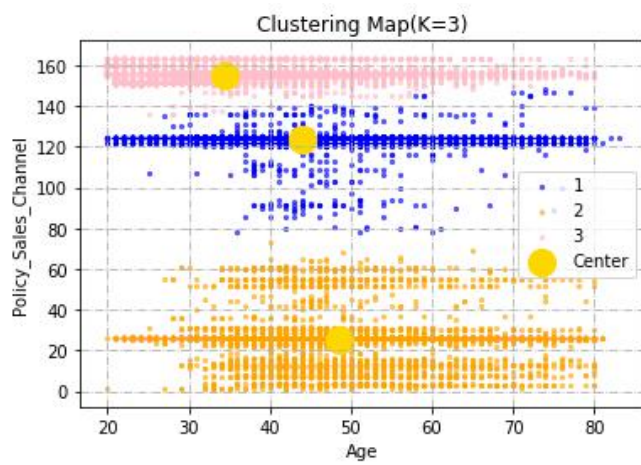
Clustering:

Test1: Clustering for 'Age' & 'Region Code'



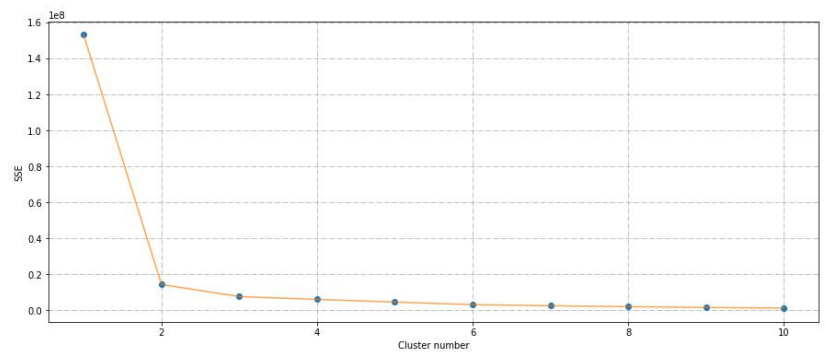
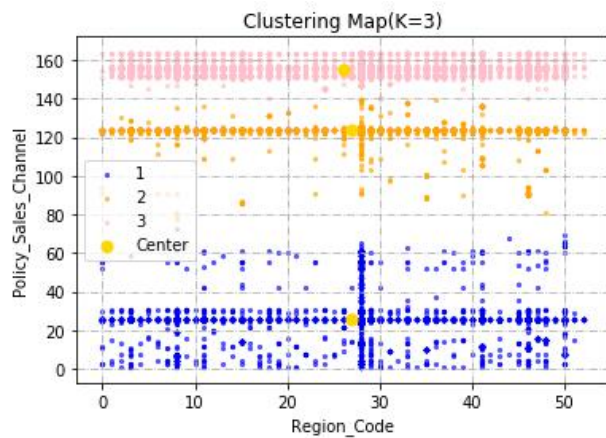
(Figure 14)

Test2: Clustering for 'Age' & 'Policy Sales Channel'



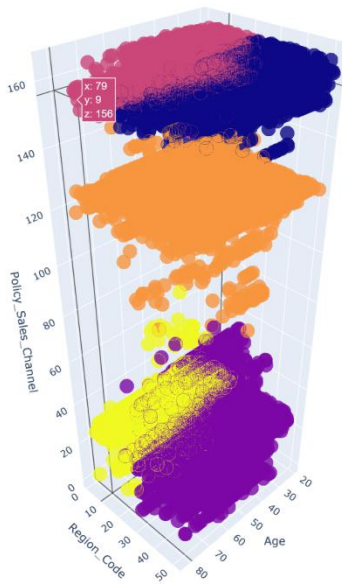
(Figure 15)

Test3: Clustering for 'Region Code' & 'Policy Sales Channel'



(Figure 16)

Test4: Visualization for 'Age' & 'Region Code' & 'Policy Sales Channel'



(Figure 17)

Testing Analysis:

Test1 forms a great division

Test2&Test3 performs not that good. 'Policy Sales Channel' is actually a multipolar dominant attribute because channel 24,124 are relatively more successful in selling insurance.

Customer Analysis:

Insurance companies can divide customers into "Regional age groups", that is, when the insurance company sells products in a certain region or a certain state, it can focus on customers of corresponding age groups in this region, who are likely to be the most likely to want to buy auto insurance in this region. This will greatly reduce their target customers, find the customers who are most likely to buy the car insurance, thus saving the company time and labor to find the target customers.

III. Conclusion

Finally, we completed the algorithm, obtained the data visualization results, the establishment

of the classifier and deviation calculation and correction. Using these algorithms and data, we can solve the original problem of helping companies increase revenue.

Through the visual table, we can see that the customers who are willing to buy automobile insurance have some characteristics. For example, all the customers who buy vehicle insurance are people with driving licenses, and those with the vehicle age of 1 to 2 years are more likely to buy car insurance; the customer's vehicle has been damaged will want to buy vehicle insurance, and no damage has no purchase intention. Through the results of the classifier and the fitting of logistic regression, we can build a prediction model with high accuracy to predict whether customers are willing to buy automobile insurance. If an insurance company wants to increase the sales rate of its insurance, it can call all the existing customer information of the company, substitute all the information into the model, generate the prediction results, and get the list of customers who are willing to buy auto insurance.

In addition, insurance companies can arrange salesmen who are specially responsible for such customers, which will greatly improve the success rate of promotion. For example, salesperson A is more suitable to sell products to women, then salesperson A will be arranged to sell car insurance to female customers who own cars; salesman B is more suitable to sell products to the elderly, then salesman B will be arranged to sell car insurance to customers who are over 60 years old and own a car.....

In this way, the company can greatly improve the sales rate of their car insurance, so as to avoid wasting employees' time by blindly introducing their car insurance to every customer, so that they can expand sales and reduce costs, which greatly increases the profit of the insurance company.