INSURANCE CROSS SELLS PREDICTION

BY WAIWIT CHAROENWILATPONG

CONTENT

- Objective
- Descriptive Analytics
- Diagnostics Analytics
- Predictive Analytics
- Business use case

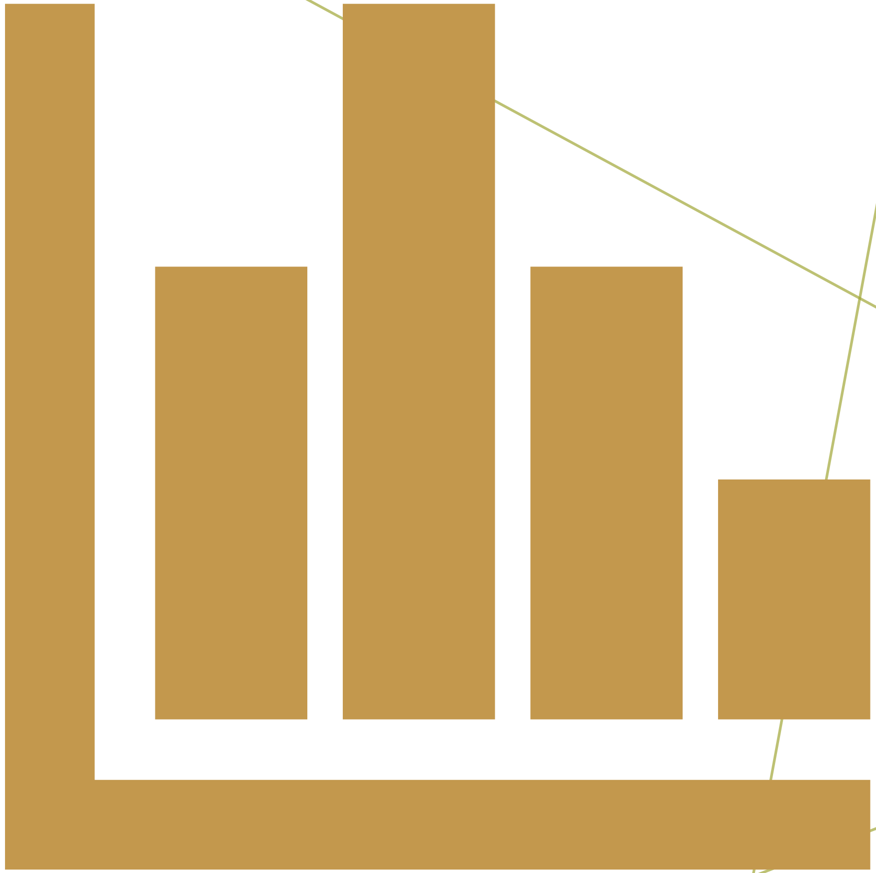
IMPLEMENTATION

OBJECTIVE

OBJECTIVE

Insurance Cross Sales Prediction is algorithm which is aiming to find Health Insurance Customers who has potential to purchase Vehicle Insurance.

DESCRIPTIVE ANALYTICS



DATASET INFORMATION

- Simulation Data source: https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction
- 381,109 records for model development (from train.csv)
  - 224,999 records for training
  - 156,110 records for testing
- 127,037 records for use cases explanation (from test.csv)

DATASET INFORMATION (CONTINUE)

Column Name	Description	Alias
id	Customer Unique ID	id
gender	Customer Gender	gender
age	Customer Age	age
driving_license	Obtaining Driving License (0: Doesn't Have, 1: Have)	license
region_code	Unique code for Region of Customer	region
previously_insured	Obtaining Vehicle Insurance (0: Doesn't Have, 1: Have)	prev
vehicle_age	Age of Vehicle	v_age
vehicle_damage	Customer's Vehicle got damage in the past (0: Didn't get damage, 1: Got damage)	v_dmg
annual_premium	Amount to pay as premium in a year	prem
policy_sales_channel	Anonymized code of channel outreached customer	channel
vintage	Number of day which customer associate with company	vintage
response	Offer Response (0: Not interest, 1: Interest)	resp

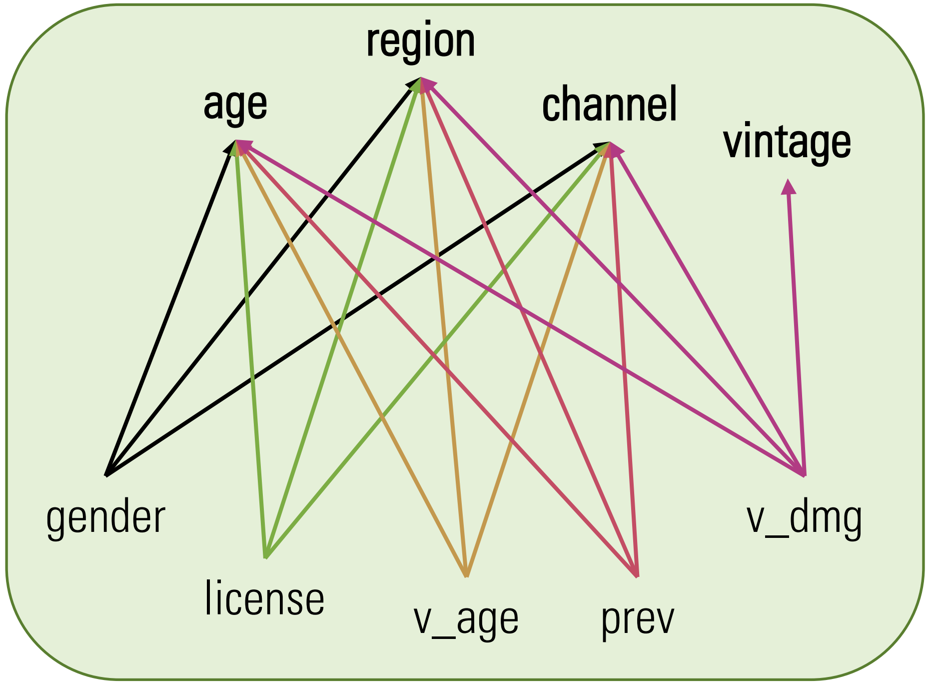
DIAGNOSTICS ANALYTICS

DEPENDENCE BETWEEN CATEGORICAL VARIABLES

Chi-Square testing shows dependency between categorical variables are presented as chart.

Customer age, region, and sales channel are not independent with customer gender, having license, vehicle age, having insurance, and having vehicle damage record. While vintage is not independent with having insurance and having vehicle damage record only.

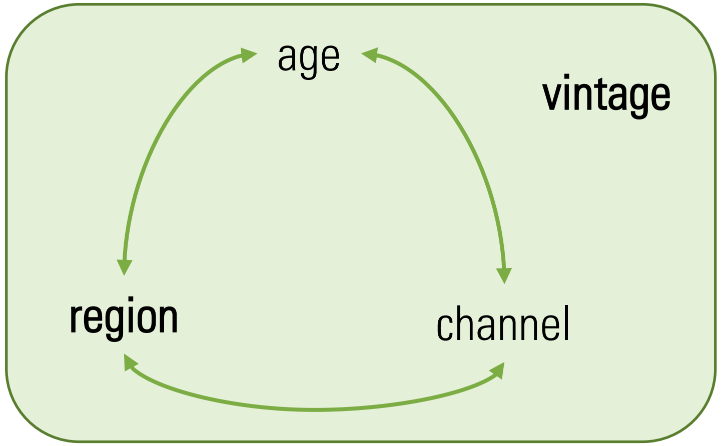
So, we should recheck dependency between age, region, sales channel, and vintage



DEPENDENCE BETWEEN OUTCOME AND CATEGORY VARIABLES (CONTINUE)

Chart shows dependence between *response, age, region, channel*, and *vintage.* however, vintage is independent to others.

This means population can be distinguished in the same manner between *response, age, region,* and *channel;* but differently from *vintage*.



ANNUAL PREMIUM VS CATEGORY VARIABLES (INNER CIRCLE)

These are ANOVA testing results between annual premium and *age, region,* and *channel.*

Even though, p-value shows all variables have potential to distinguish population by mean of annual premium, **but F-value indicated *region* is the best among them.**

One-way ANOVA No split

Compare mean of Annual_Premium for each Age

Age	Count	Mean
24	14981	30206.110073
23	13963	29864.743393
22	12240	29983.673039
25	11779	29736.040071
21	9669	29574.819733
26	7905	28502.565465

Hypothesis	
Tested hypothesis	Annual_Premium means are identical in all populations
Significance level	0.05
Results	
F-value	38.433706478
p-value	0

One-way ANOVA No split

Compare mean of Annual_Premium for each Region_Code

Region_Code	Count	Mean
28.0	59071	36116.046402
8.0	18974	34292.711289
46.0	11587	26804.264434
41.0	10568	30113.307627
15.0	7843	29115.7121
30.0	7200	25509.934444

Hypothesis	
Tested hypothesis	Annual_Premium means are identical in all populations
Significance level	0.05
Results	
F-value	1043.1590355
p-value	0

One-way ANOVA No split

Compare mean of Annual_Premium for each Policy_Sales_Channel

Policy_Sales_Channel	Count	Mean
152.0	78054	29996.792887
26.0	45058	31754.337876
124.0	42540	30425.238223
160.0	12719	24866.036009
156.0	6354	10782.480642
122.0	5465	33931.559012
157.0	3842	13235.127538

Hypothesis	
Tested hypothesis	Annual_Premium means are identical in all populations
Significance level	0.05
Results	
F-value	184.6047658
p-value	0

ANNUAL PREMIUM VS CATEGORY VARIABLES (VINTAGE)

ANOVA testing results between annual premium and *vintage* reveals that <u>mean of premium is the same among population</u> even being grouped by vintage.

One-way ANOVA No split

Compare mean of Annual_Premium for each Vintage

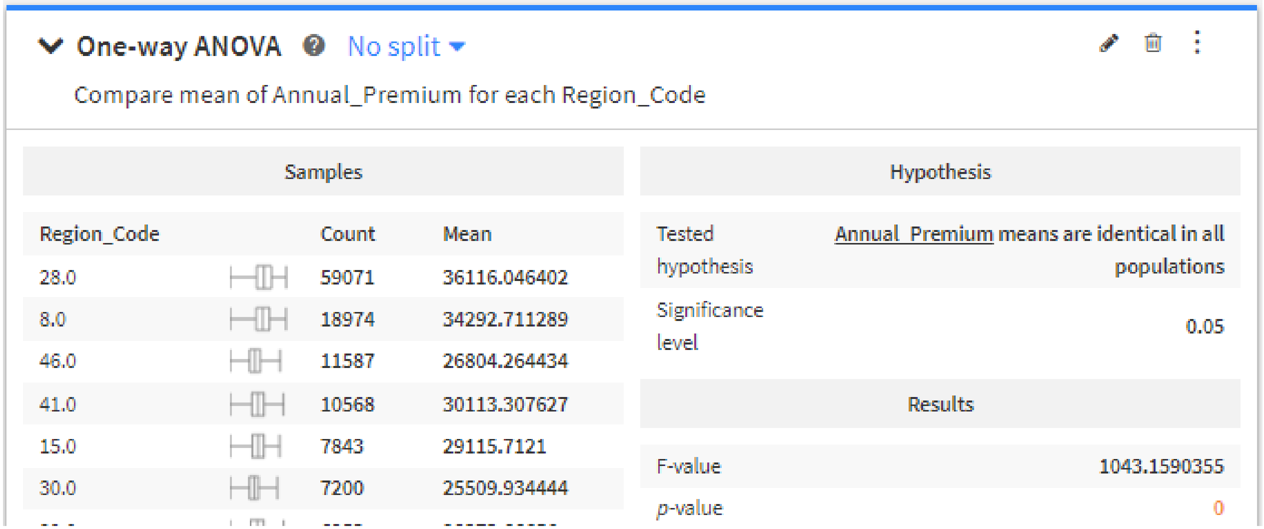
Samples		
Vintage	Count	Mean
92	835	28761.488623
249	829	29762.93848
248	824	29536.36165
63	818	28789.611247
160	817	29001.849449
56	814	28904.400491
73	814	29217.739558

Hypothesis	
Tested hypothesis	<u>Annual_Premium</u> means are identical in all populations
Significance level	0.05

Results	
F-value	1.0087353651
p-value	0.4473767972

ANNUAL PREMIUM VS CATEGORY VARIABLES SUMMARY

ANOVA testing results are indicated that combination between annual premium and region of customer might have potential to reveal difference of customer’s interest.



One-way ANOVA No split

Compare mean of Annual_Premium for each Region_Code

Samples

Region_Code	Count	Mean
28.0	59071	36116.046402
8.0	18974	34292.711289
46.0	11587	26804.264434
41.0	10568	30113.307627
15.0	7843	29115.7121
30.0	7200	25509.934444

Hypothesis

Tested hypothesis	Annual_Premium means are identical in all populations
Significance level	0.05

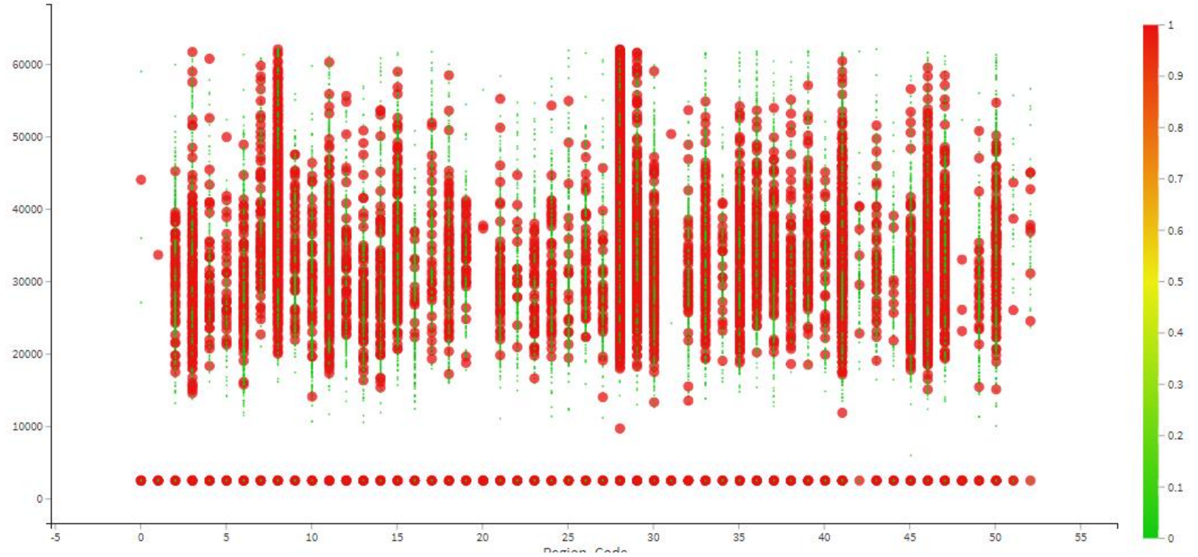
Results

F-value	1043.1590355
p-value	0

VISUALIZATION WITH CUSTOMER RESPONSE

In the chart, we can see patterns of customer, who is probably interested by condition of region and premium amount.

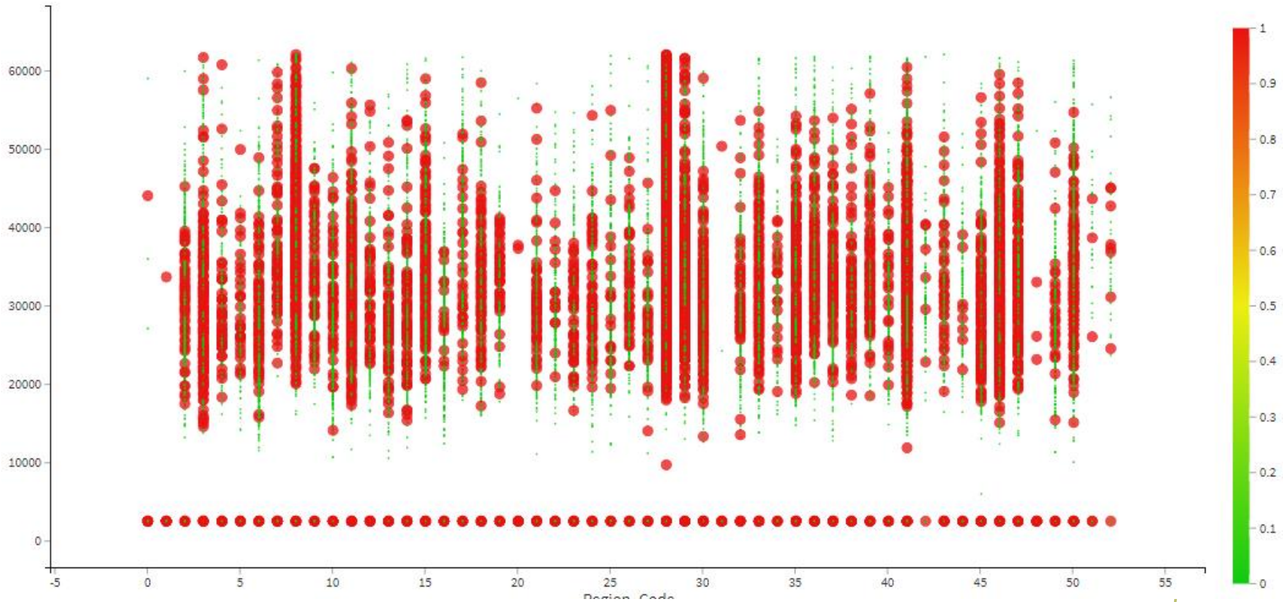
So, it is possible to use this pattern to determine possibility of interest in vehicle insurance.



PREDICTIVE ANALYTICS

PREDICTION MODEL CANDIDATES

From this chart, it is obvious that **conditions of customer's interest is based-on ranges of premium amount within each region.** So, **Random Forest** might work well on this case. However, amount of premium might have variance; therefore, **Logistic Regression** might be considered as option.



MODEL TRAINING RESULT

The training result shows performance of each model, plus Ensemble for alternative. It seems all models work equally.

Model	AUC
Random Forest	61.4%
Logistic Regression	61.3%
Ensemble	61.7%

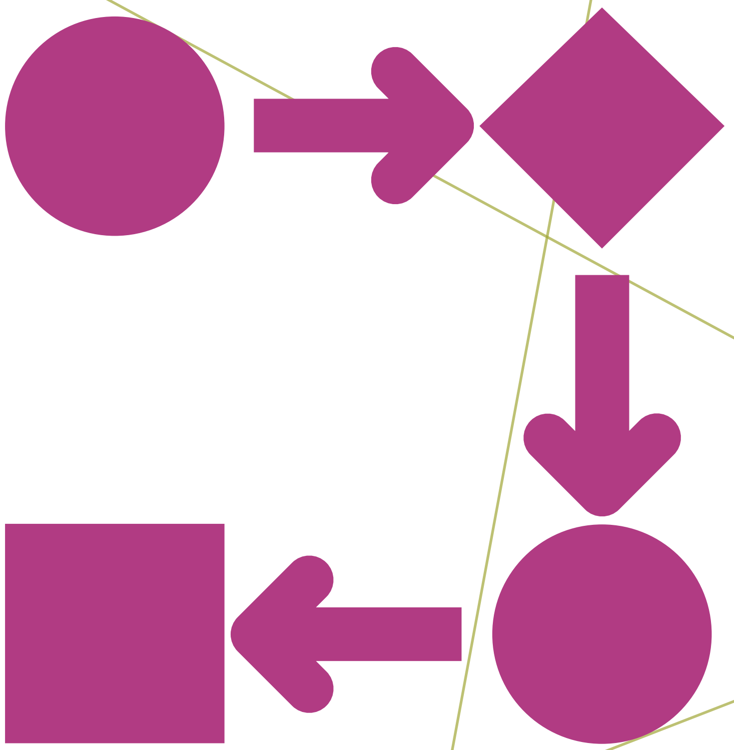
MODEL TESTING RESULT

From the testing results, It shows that models work better slightly if it has regression part.

Therefore, we will continue with Ensemble Model.

Model	AUC
Random Forest	61.4%
Logistic Regression	61.7%
Ensemble	61.9%

BUSINESS USE CASE EXAMPLE

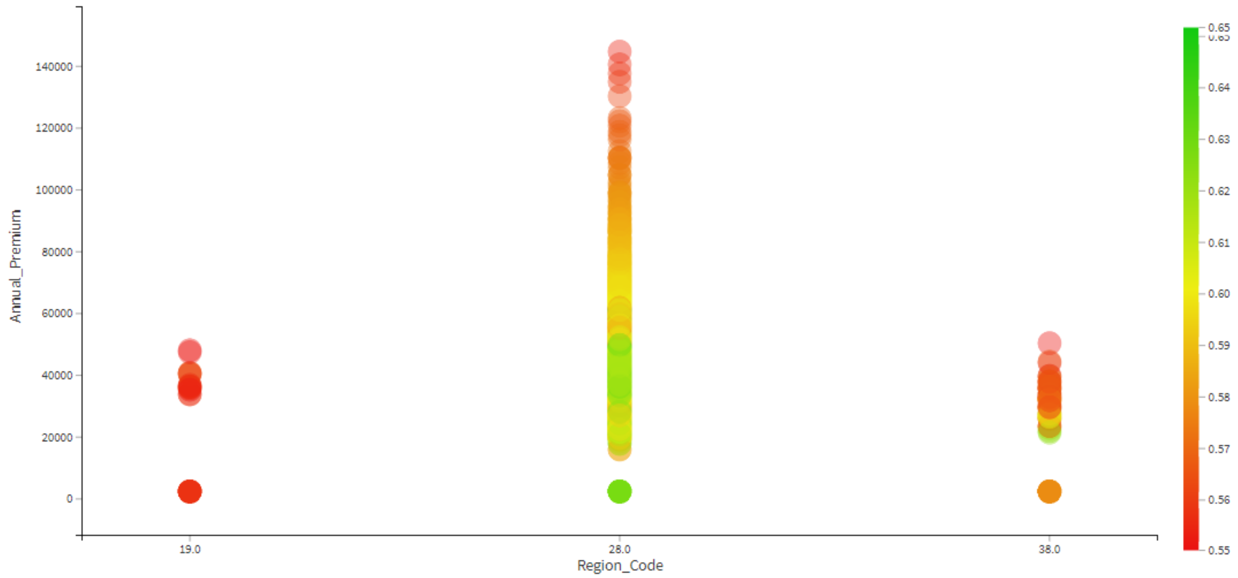


CROSS SELLS STRATEGY

This is an example that we use the predict models.

Once we get prediction result from model, we can group customer to 3 groups as table.

With additional information such as, region and amount, we can see “Region 28.0” has high potential to sell vehicle insurance.



Category	Probability	# Customers	Top Region Code
Hi-Potential	60% +	26,642	28.0 (26,617), 38.0 (25)
Lo-Potential	55 – 60 %	9,368	28.0 (8,777), 38.0 (416), 19.0 (175)
Poor Interest	50 – 60 %	6,252	-
Ignore	Low Probability	84,775	-
Total		127,037	28.0 (35,394), 38.0 (441), 19.0 (175)

THANK YOU

By Waiwit Charoenwilatpong 6410414003