Indrajith Senevirathne
September 2021

ML Analysis and Prediction of the E – Marketing Sales Conversion and Lead Efficiency

Overview

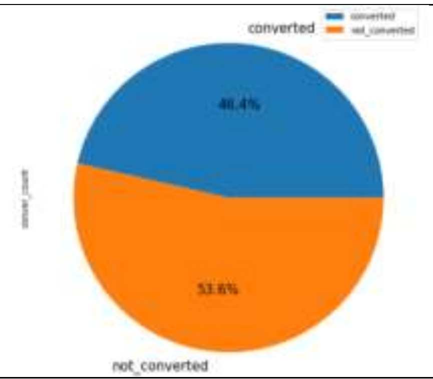
Analyzed data sets include:

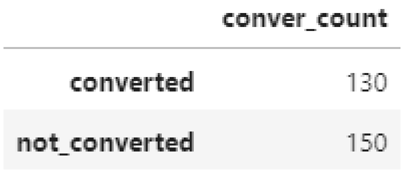
a. Constructed sales conversion data set (conversion_rates.csv) to investigate the amount of conversions at the onset of new product introduction.
b. Constructed lead and sales conversions data sets (lead_sales_stats.csv, names_id_age.csv) to analyse and assess the lead efficiency

Summary of the analysis findings follows

Part 1. Sales Conversions

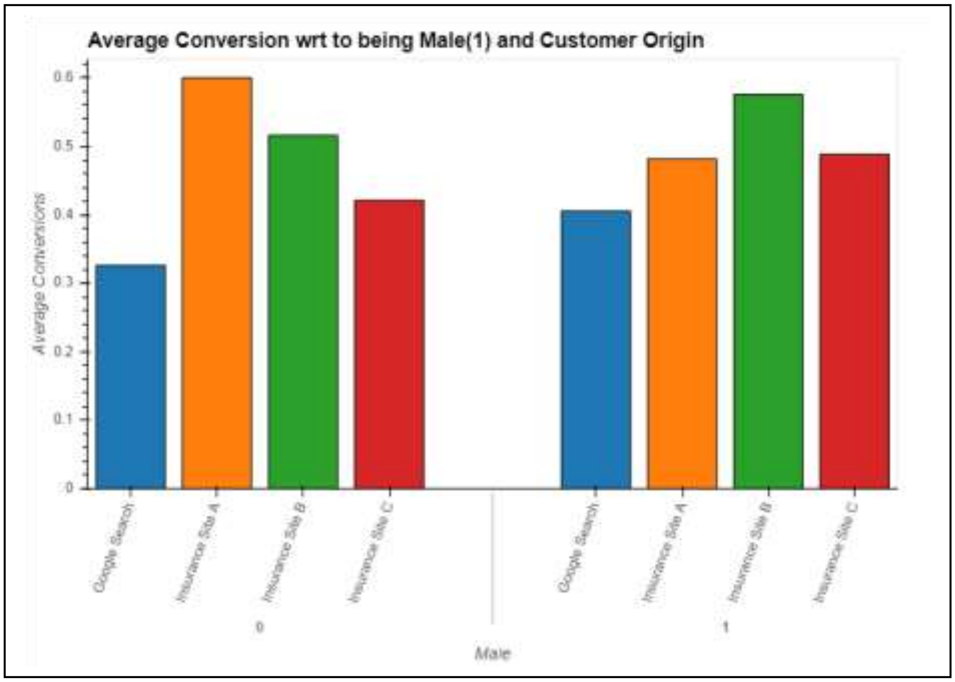
1. The *new product* has a GOOD conversion rate: conversions = 46.4% as it is much higher than typical e-commerce conversion rates (highest e-commerce conversion rates are for Health & Welbeing Category ~4% while Amazon boasts a conversion rate of 13%). However, it is not possible to assess if the new product has improved the conversion rate unless we have the conversion rate of the previous product. [see below]

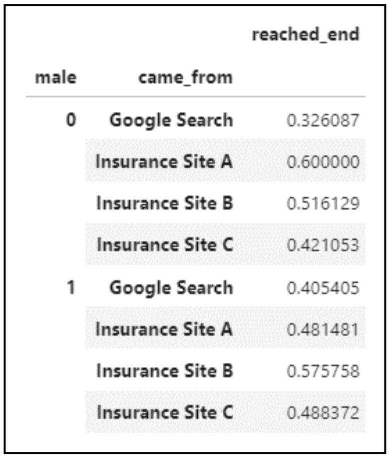




	conver_count
converted	130
not_converted	150

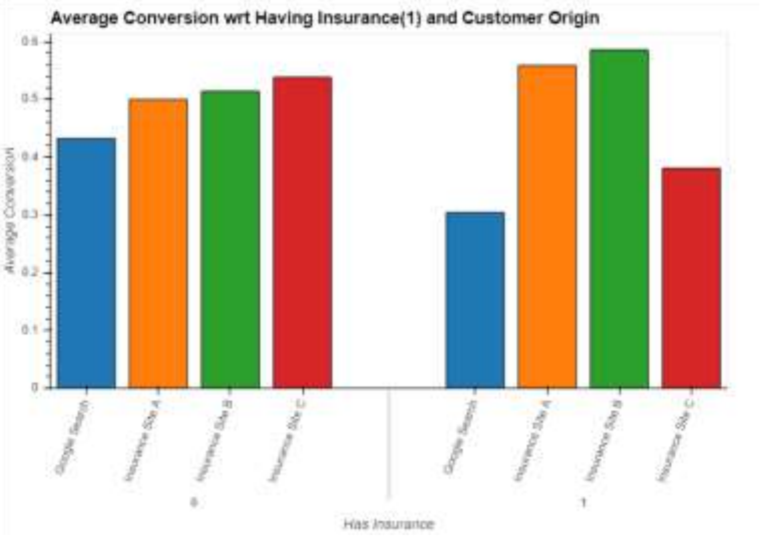
2. Product Improvement suggestions:
   a. If possible product should be targeted more towards the female demography in Insurance site A while it should be targeted more towards male demography in the Insurance site B. Google leads are not dependable as their conversions are the worst for both male and female demographics. [see below]

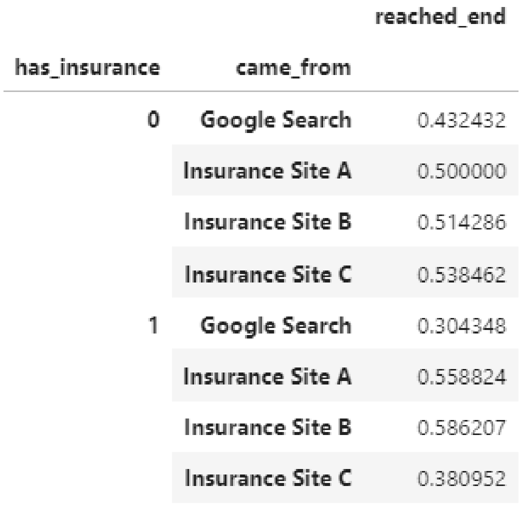




		reached_end
male	came_from	
0	Google Search	0.326087
	Insurance Site A	0.600000
	Insurance Site B	0.516129
	Insurance Site C	0.421053
1	Google Search	0.405405
	Insurance Site A	0.481481
	Insurance Site B	0.575758
	Insurance Site C	0.488372

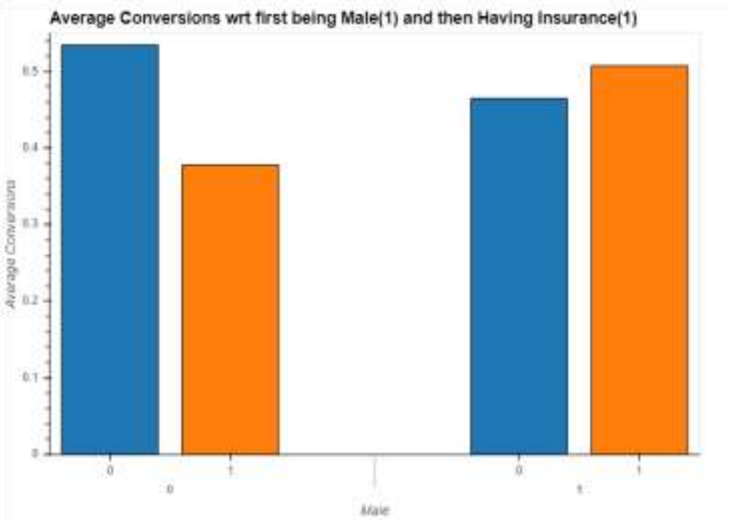
b. If possible product should be targeted more towards the individuals without insurance in Insurance site C while the product should be targeted for individuals with insurance in Insurance site B. Google leads are not dependable as their conversions are the worst for both individuals with or without insurance. [see below]

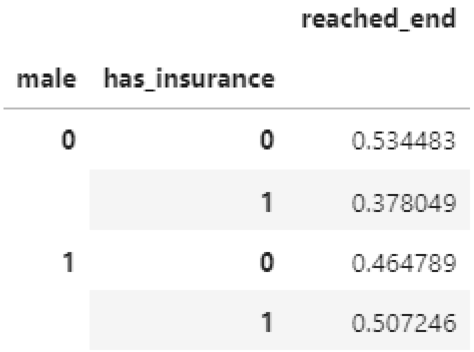




		reached_end
has_insurance	**came_from**	
0	**Google Search**	0.432432
	Insurance Site A	0.500000
	Insurance Site B	0.514286
	Insurance Site C	0.538462
1	**Google Search**	0.304348
	Insurance Site A	0.558824
	Insurance Site B	0.586207
	Insurance Site C	0.380952

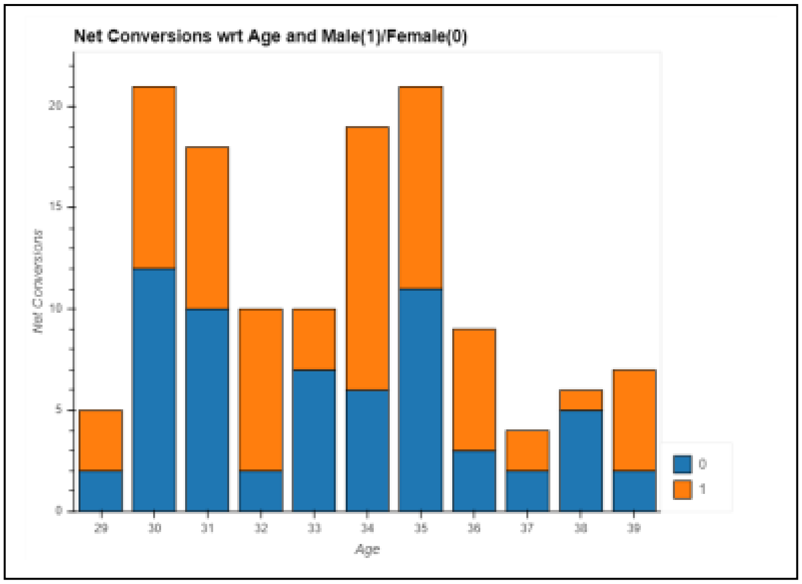
c. Advertising campaigns should be targeted heavily on the females without insurance and males with insurance as these cohorts respond well to the product. [see below]



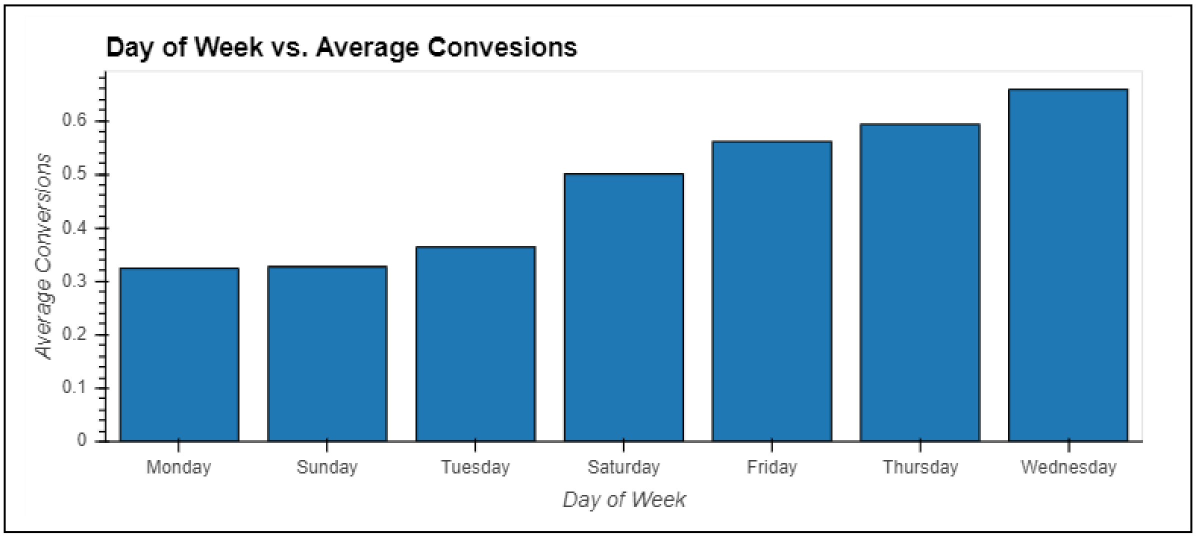


		reached_end
male	**has_insurance**	
0	**0**	0.534483
	1	0.378049
1	**0**	0.464789
	1	0.507246

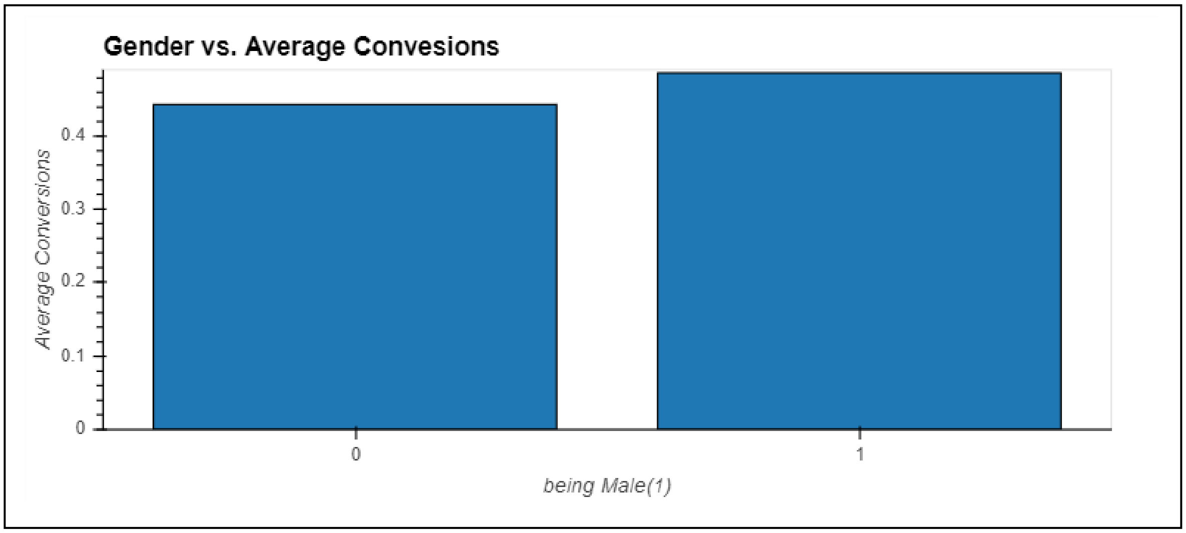
d. Conversion rates peak at the age of 30 years (primarily female) and again at 34-35 years (male and female both). Advertisements should target these two demographics, and the product should also cater to these cohorts, more likely increasing conversion rate further. [see below]



e. On average, the highest number of conversions occur on Wednesdays, whereas the lowest conversions occur on Mondays. Therefore, the advertising campaigns will be more effective on Wednesdays. (Assumption: the data set is comparable to any other such data set taken at a different set of days)



f. On average conversion rate for male demography is higher; this is relevant because the net demography of individuals is equally split between 50% (male) and 50%(female).



Predicting if an individual is likely to convert at the end of the line towards the new product using ML

A predictive classification ML model was used to model the conversions: 1 being conversions and 0 being non-conversions. This sample size is (280) extremely small for ML models; as such, this can only be considered as an exercise of feasibility.

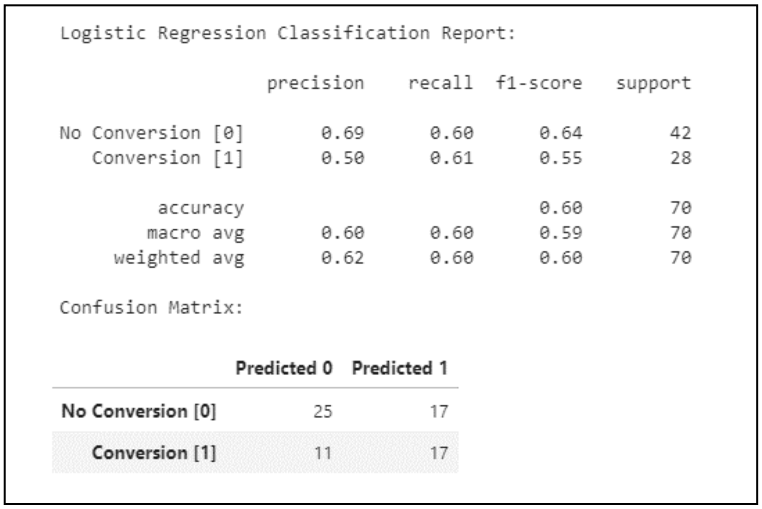
Data were seperated to train (75%) and test (25%) using Scikit-Learn, train_test_split. Next they were scaled using StandardScaler.

It is evident that the division of my small sample didn't help as the amount of training data (210) got even smaller, and the test data was also small (70). Model deployed is a LogisticRegression(LR) with solver at SAGA (a variant of Stochastic Average Gradient SAG). Predicted values were obtained and compared with test (y_test) values. LR classification report was obtained, the accuracy score is a poor 60%. Also, the Precision and Recall scores were ~60%. Conversions have a low Precision score of 50%, as is only 1 in 2 predicted to convert will Truly convert.

These metrics are seen on the Classification Report and Confusion Matrix.

As is the model has very low predictability and likely can be improved with a sufficiently large data set to train.

[see below]

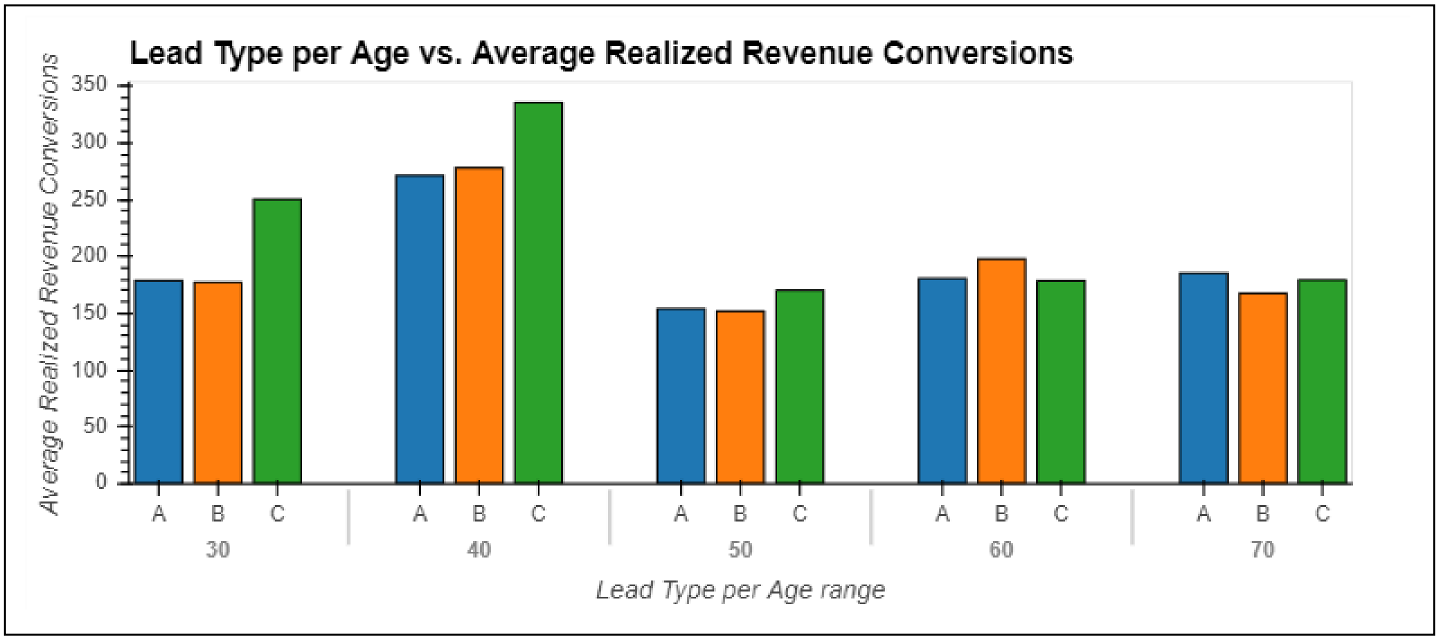


```
Logistic Regression Classification Report:

                    precision    recall  f1-score   support

 No Conversion [0]       0.69      0.60      0.64        42
    Conversion [1]       0.50      0.61      0.55        28

          accuracy                           0.60        70
         macro avg       0.60      0.60      0.59        70
      weighted avg       0.62      0.60      0.60        70

Confusion Matrix:
```

	Predicted 0	Predicted 1
No Conversion [0]	25	17
Conversion [1]	11	17

Part 2. Lead Efficiency

Conversion rates are the highest for the 35-45 year demography. The Intent Lead algorithm is not trustworthy; C (supposed to be the least intent) generates on average more realized conversions in 3 out of 5 cohorts. [see below]


Lead Type per Age vs. Average Realized Revenue Conversions
Average Realized Revenue Conversions
A
B
C
Lead Type per Age range