ASSIGNMENT 3: DATA ANALYSIS PROJECT FOR MARKETING CAMPAIGNS

Simon Lim

24661225

University of Technology Sydney

36103 Statistical Thinking for Data Science

Dr. Alice Dong

1 INTRODUCTION

A telecommunication company launched a marketing campaign to improve the adoption of their new subscription plan among customers. The telecommunication company has recorded various client's information, including age, job, education, housing, loan, last contact duration, number of contacts performed before the campaign and consumer price index etc. The company requested assistance in gaining a comprehensive understanding of their customers and identifying the customer segments that show the likelihood of the highest responsiveness to marketing campaigns.

In this regard, the main objective of the project is to build statistical learning models that can successfully identify the customer segments that are the most responsive to marketing campaigns. Furthermore, the project also aims at finding three effective business strategies that can help the company in making decisions for their marketing campaigns. By achieving those two goals, the company can gain a deeper understanding of their customers and launch a more effective campaigns, in order to promote the adoption of new subscription plan.

2 METHODS

2.1. Data Understanding/Preparation

A dataset contained 41180 rows and 21 columns with a response variable 'y' indicating whether customers previously subscribed to a plan. The response variable was imbalanced with 36531 of no subscription and 4637 of subscription. Also, the data type of all columns was object with a couple of columns containing numeric values. Since some of columns consisted of numeric values, their data type needed to be converted into integer of float in later stage. While there were no null values in the dataset, there were 12 duplicates, which were removed for improving data quality.

2.2. Data Conversion

The values in 'y', the response variable, were replaced by numeric values (i.e., from 'no' and 'yes' to 0 and 1) and data type was also converted to an integer. Additionally, data types of

‘age’, ‘campaign’, ‘duration’, ‘pdays’ and ‘previous’ columns were converted from object to integer and data types of ‘emp.var.rate’, ‘cons.price.idx’, ‘cons.conf.idx’, ‘euribor3m’ and ‘nr.employed’ columns were converted to float type. Lastly, the rest of data in columns (‘job’, ‘marital’, ‘education’, ‘default’, ‘housing’. ‘loan’, ‘contact’, ‘month’, ‘day_of_week’ and ‘poutcome’) were converted into numeric values using ‘OrdinalEncoder()’, which enabled to convert categorical values into ordinal numeric values.

2.3. Outliers

There was one column ‘pdays’, which contained a unique value of 999. Given that ‘pdays’ column represented number of days that passed by after the client was last contacted from a previous campaign and was also an outlier from other unique values, I decided to change a unique value of 999 to 0 for improving data quality and modelling.

2.4. Bayesian Estimation

Bayesian estimation is used to estimate the relative effects of parameters on the model, along with their corresponding 95% credible intervals. The rationale behind choosing Bayesian estimation over maximum likelihood estimation is that because prior information was available on some columns, such as ‘duration’ (last contact duration), ‘previous’ (number of contacts performed before the campaign) and ‘y’ (whether subscribed or not previously). Markov Chain Monte Carlo (MCMC) sampling was used to estimate posterior distributions in Bayesian inference, along with 100000 number of steps.

```
Parameter Estimates:
age: 0.02230760051416149 (95% Credible Interval: -0.004899396271574385, 0.04915151099590106)
job: 0.0030341405307153855 (95% Credible Interval: -0.021441520336838223, 0.028501486014849618)
marital: 0.01889223845752003 (95% Credible Interval: -0.006849486983687074, 0.045032797422483825)
education: 0.03924081633076967 (95% Credible Interval: 0.013954952435710807, 0.0640434541597552)
default: -0.025977877936528386 (95% Credible Interval: -0.05077205566596078, -6.162637462529819e-05)
housing: 0.003163362940091303 (95% Credible Interval: -0.020875070487173387, 0.02680106428129877)
loan: 0.005833460166797301 (95% Credible Interval: -0.01786757286245734, 0.03028421877203808)
contact: -0.18029027567116360 (95% Credible Interval: -0.21732959892572398, -0.14334979839855716)
month: -0.13370383520831083 (95% Credible Interval: -0.16695242103960453, -0.09826636056812652)
day_of_week: 0.01602349635930621 (95% Credible Interval: -0.007761027423819452, 0.03983080043281816)
duration: 0.6407340910729753 (95% Credible Interval: 0.60867510099718, 0.6719682444117766)
campaign: 0.025071624227701178 (95% Credible Interval: 0.0001538504737600456, 0.04889251725209906)
pdays: 0.18219943548924958 (95% Credible Interval: 0.07709667199759665, 0.3022816254656743)
previous: 0.48239259600983636 (95% Credible Interval: 0.3824017669316538, 0.5803693536327835)
poutcome: 0.49082508283902326 (95% Credible Interval: 0.4102357889024525, 0.5694582958276831)
emp.var.rate: -0.9240686702274085 (95% Credible Interval: -1.085150397000511, -0.7575450329487448)
cons.price.idx: 0.3671162765167177 (95% Credible Interval: 0.29032546059026193, 0.4480503876069694)
cons.conf.idx: 0.10291696421164723 (95% Credible Interval: 0.054210321258773224, 0.15456140101320798)
euribor3m: 0.7143176015411603 (95% Credible Interval: 0.4527592121064448, 0.9597986425627832)
nr.employed: -0.43804403908378725 (95% Credible Interval: -0.6062818936613785, -0.26015466697793177)
```

Figure 1. Bayesian Estimation Using MCMC (Markov Chain Monte Carlo) sampling.

Accordingly, 'duration' (coeff = 0.641) had a strong positive effect on the target variable and 'Emp.var.rate' (coeff = -0.924) had a strong negative effect on the target variable. Additionally, 'Poutcome' (coeff = 0.491)', 'Previous' (coeff = 0.482) and 'Euribor3m' (coeff = 0.714) also showed a significant positive effect on the target variable.

2.5. Modelling

One parametric model (Logistic Regression) and two non-parametric models (LightGBM and SVM) were established and assessed to predict the success of a marketing campaign for a given customer in binary classification.

Logistic Regression (Parametric Model): The rationale behind choosing logistic regression is that the response variable is binary and also linearly separable. Logistic regression is powerful algorithm in binary classification when data is linearly separable.

LightGBM (Non-parametric Model): LGBM is accurate and well-suited for a large dataset. Given that the dataset contained 41180 data, LGBM is likely to result in optimal performance in binary classification analysis.

3 RESULTS

3.1. Logistic Regression (Parametric Model)

Two kinds of logistic regression models were trained, predicted and assessed, one with default logistic regression and the other with Bayesian optimization. The models performed very well in predicting negative class (f1-score = 0.95, AUROC = 0.70) but relatively performed poorer in identifying positive class (f1-score = 0.52, AUROC = 0.70).

	precision	recall	f1-score	support
Less likley to subscribe	0.93	0.98	0.95	7318
More likely to subscribe	0.68	0.42	0.52	916
accuracy			0.91	8234
macro avg	0.81	0.70	0.73	8234
weighted avg	0.90	0.91	0.90	8234

Figure 2. The table of logistic regression performance score

3.2. LightGBM Classifier (Non-parametric Model)

Similar to logistic regression models, default model and Bayesian optimization model were trained and tested for LGBM classifier. The models showed stronger performance than logistic regression in terms of F1 score and AUROC. However, models still performed poorer in identifying positive class (f1-score = 0.61, AUROC = 0.76) than negative class (f1-score = 0.96, AUROC = 0.76).

	precision	recall	f1-score	support
Less likley to subscribe	0.95	0.96	0.96	7318
More likely to subscribe	0.67	0.57	0.61	916
accuracy			0.92	8234
macro avg	0.81	0.77	0.78	8234
weighted avg	0.92	0.92	0.92	8234

Figure 3. The table of LGBM performance score

4 DISCUSSION

F1 score and AUROC score were used as main metrics to evaluate the performance of models. F1 score and AUROC score are both optimal performance metric especially when data are imbalanced. Given that the target variable 'y' is imbalanced with 36531 of negative class and 4637 of positive class, those two metrics are suitable as performance metric. Also, positive class (i.e., subscribed to a plan) is crucial in identifying customer segments, who are likely to respond. Since F1 score particularly focuses on the model's ability to correctly identify the positive class, it is a suitable metric performance.

LGBM models showed a stronger performance than logistic regression models in identifying customers, who are the most likely to respond to the marketing campaign. However, given that positive and negative classes were imbalanced, it was challenging to obtain a high f1-score. Nevertheless, the highest f1-score obtained in identifying positive class was **61%**, which is a moderate performance score.

A few effective strategies could be extracted through Bayesian estimation. First, there was a strong negative correlation between employment variation rate and subscription behavior. That is, as the employment variation rate of clients is high, clients are less likely to respond to the campaign. Hence, it would be suggested to target customer segments that have low employment variation rate. Furthermore, there was also a strong positive correlation between last contact duration and number of contacts and subscription behavior. It is suggested that if clients have had a number of contacts for considerable duration, they are more likely to respond to a marketing campaign. Lastly, there was a strong positive correlation with EURIBOR-3 month., which is the average interest rate at which a selection of European banks lends funds to one another in the interbank market. Variations in this rate can considerably influence the customer's responsiveness to the campaign. As this rate is high, clients are more likely to respond to the campaign.

5 CONCLUSION

In conclusion, the project aimed at establishing statistical learning models that predict the success of the campaign for a given customer and extracting three effective strategies to

help the company in making decisions for their campaigns. The models successfully predicted customer segments, who are the most likely to respond to marketing campaign and Bayesian estimation enabled to extract effective strategies for the campaign.

6 REFERENCE

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning : with applications in R (SECOND). Springer.

Herbst, E.P. and Schorfheide, F. (2015) *Bayesian estimation of DSGE models* [Preprint]. doi:10.23943/princeton/9780691161082.001.0001.

7 APPENDIX

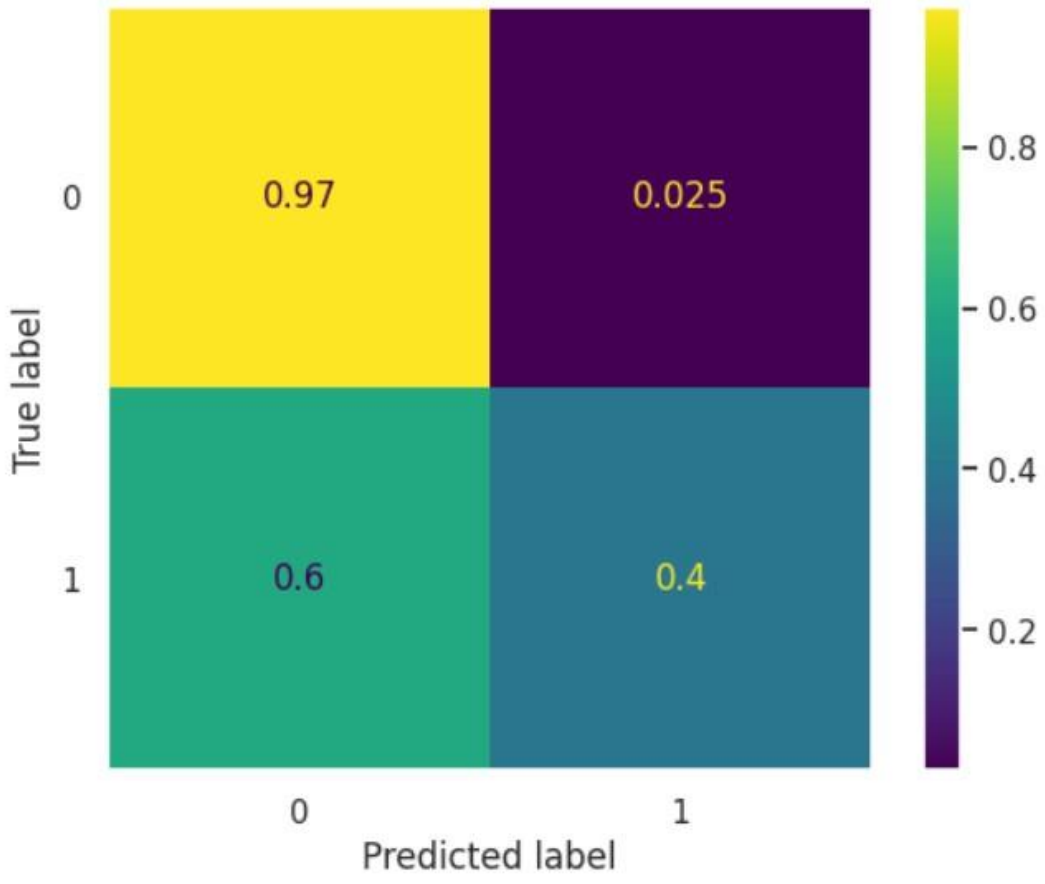

Figure 4. The Confusion Matric of Logistic Regression Model

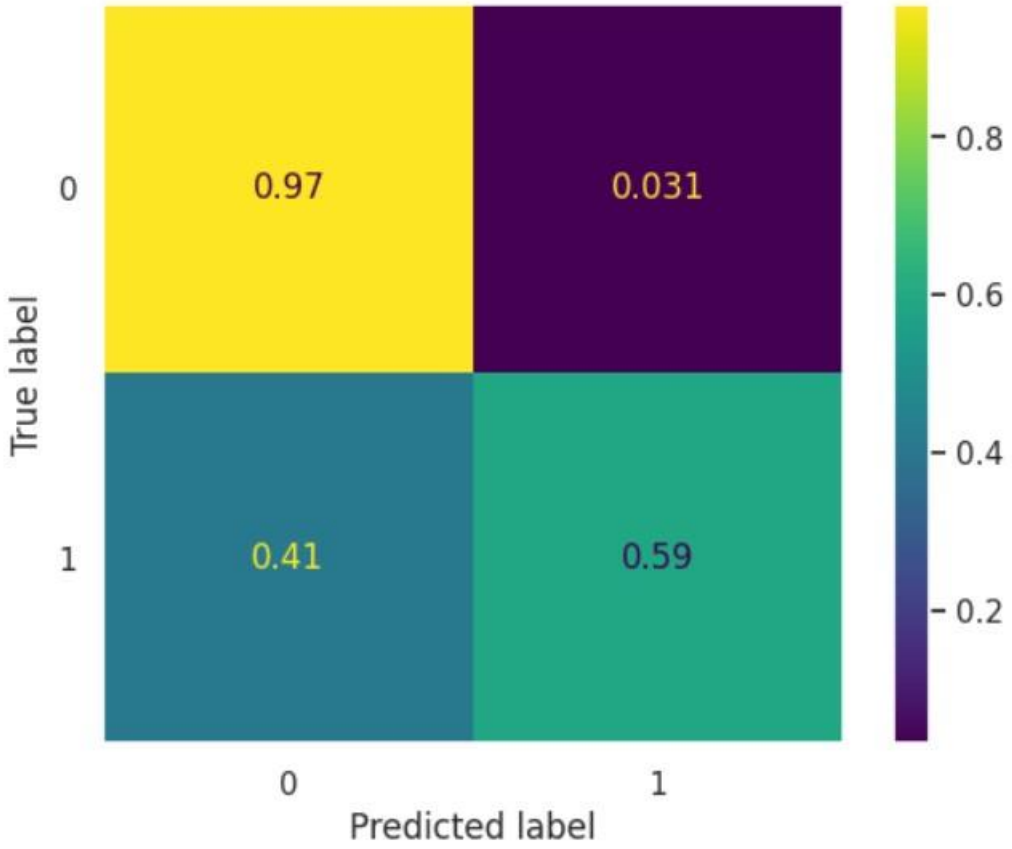

Figure 5. The Confusion Matric of LGBM Model

```
best_params
```

```
OrderedDict([('C', 0.05680227153965569),
             ('max_iter', 100),
             ('penalty', 'l2'),
             ('solver', 'liblinear')])
```

```
log_reg_opt = LogisticRegression(**best_params)
log_reg_opt.fit(X_train, y_train)
```

```
LogisticRegression
LogisticRegression(C=0.05680227153965569, solver='liblinear')
```

Figure 6. Best Parameters for Logistic Regression using Bayesian Optimization

```
best_params
```

```
OrderedDict([('colsample_bytree', 0.9719327800362794),
             ('learning_rate', 0.1066075305512463),
             ('min_child_samples', 63),
             ('num_leaves', 12),
             ('reg_alpha', 0.28656260463532574),
             ('reg_lambda', 0.2693303804120745),
             ('subsample', 0.8400204481592152)])
```

Figure 7. Best Parameters for LGBM using Bayesian Optimization