Survey Incentive Model for Email

By Chen song, Data Scientist

loxz digital

Table of Contents

ABSTRACT

This report describes the survey incentive model developed by Loxz Digital. Currently, this model provides RealTimeML predictive analytics on the engagement rate of survey incentive campaigns with incentives built within the workflow of an email campaign before deploying it.

The survey incentive is served within milliseconds, within the workflow of an email editor, and derives the optimal recommendation incentive amount based on the target variable for the highest engagement rates including the highest click-to-open rate and conversion rate based on the selected target variable, survey type, industry, survey incentives, and survey length. The variables to be optimized, in this case, are the incentive rewards amount and survey length, and the independent variables would be industry type and survey type created. This model focuses on the survey campaign for a practical reason.

The current dataset contains 54864 samples of data with features including survey type, industry, survey incentive value,
and survey length(in minutes). The machine learning algorithm used in this model is Random Forest Regression which is an ensemble of decision trees. The model is able to provide the highest accuracy of 98.87% with our current dataset. We believe that the model can achieve a higher accuracy using transfer learning from additional data.

Ⅰ. INTRODUCTION

Survey incentives or survey rewards are used in particular in survey email campaigns to encourage users to take action on the email and complete the survey. An incentive can come in a variety of offers. It can be money or in many cases an Amazon gift card that campaign engineers provide to respondents to increase email engagement. According to research from Indiana University Bloomington, "incentivizing survey completion with a generic gift card raised response rates by 19% over no incentive, while a gift card for a more specific retailer raised rates by only 12%."(*Prince, H. (2021, December 8)*) [1]

Our survey incentive model is specifically developed to provide predictive analytics prior to deployment on the survey rewards values and survey length in email campaigns, regardless the forms of the survey incentives. The model provides predictions on an email campaign for the selected email engagement target metric as well as recommendations on how to optimize the survey rewards and survey length in the email in order to maximize the targeted campaign metric. To help email campaign engineers better run the email campaigns, we provide 2 sets of recommendations:

(1) If the email campaign engineers want to keep the survey length fixed and only want to optimize the survey rewards value, our model is able to provide recommendations on survey incentive values based on a fixed survey length.

(2) If the email campaign engineers are open to a different survey length, our model is able to provide recommendations on the optimized survey incentives as well as the survey length.

Applying Loxz Digital's survey incentive model, the users can identify in real-time, the optimum survey rewards as well the survey length to be used in the campaign to achieve the best possible email engagement rate. The campaign engineers are empowered to complete these "runs," to serve predictions within the workflow of the campaign.

II. DATA SET

Currently we trained the model on the most popular 6 types of survey:

- Human resources survey
- Customer survey
- Industry survey
- Academic evaluation survey
- Marketing survey
- Community survey

For the target variable, we utilize the 2021 Average Conversion Rate by Industry and Marketing Source from the RulerAnalytics. To fit the benchmarks in our email campaign dataset, we utilize statistical methodologies to create distributions based on the benchmarks and normalize the distributions.

Industry	No. of Survey
Academic and Education	7827
Finance and Banking	7779
Healthcare	7727
Hospitality	7812
Retail	7901
Software and Technology	7932
Professional Services	7886
Total	54864

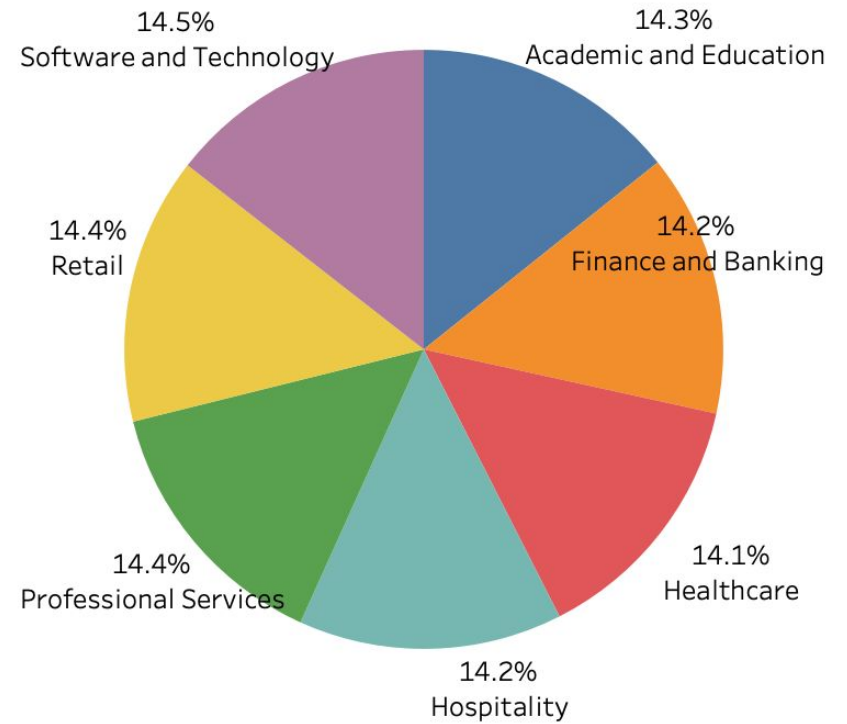**Table 1. Email distribution in the dataset across the industries**

Figure 1. Percentages of surveys distributed across industries in the dataset

III. FEATURE ENGINEERING

Because the benchmarks of conversion rate and click-to-open rate were applied in the model, we utilized statistical methodologies to create distributions to the survey sample data across different industries to get the final dependent variables. Users select the dependent variables that they want to optimize, and the model will make the prediction based on the user's selection. The benchmarks are from the Rulers Analytics website.[2] To fit the benchmarks into the training dataset, we generate "n" random numbers with a standard normal distribution (n is the number of surveys in a particular industry). Then we normalize the distributed random numbers according to the domain knowledge of experts in our team.

Along with the dependent variables, we uniformed several other features from the dataset:

- Incentive Amount: To fit the data into our case, we initially converted the discount values in the original dataset to the survey incentives. Furthermore, to make the features more flexible for the campaign engineers, we wrapped the individual incentives into ranges. Currently, the incentive amounts cover from $1 - $350 with a $10 incremental each tier.

- Survey length (minutes): Similar to the incentive amount, survey length (minutes) was also wrapped into ranges. Currently, the survey length (minutes) starts from 1 minute to 20 minutes with 5 minutes incremental for each tier. And survey that more than 20 minutes is set as an individual range.

IV. TOOL REQUIRED

A number of software and packages were used in the data transformation process and model-building process and they are all common software and packages that are frequently used in the data science community.

The implementation of the model was done in a Jupyter Notebook instance in *AWS SageMaker* using Python programming [3]. Machine learning tasks were implemented using the Scikit-learn package for Python [4]. Other commonly used Python packages are also utilized for data parsing and visualizations such as Numpy, Pandas, and ipynb. Ipywidgets is used for the demonstration and selection of targets and other parameters [5]. BOTO3 [6] is used for accessing and interacting with S3 data storage.

V. MODEL DEVELOPMENT

Predicting the best survey incentives and survey length in an email campaign with the highest engagement rate is a regression problem. So the algorithm we use in this model is a tree-based regressor: the random forest regressor.

The model predicts click-to-open rate and conversion rate which are both continuous values requiring a regression model for predictions. The regression model takes the set of features and the target variables as inputs and trains the model for predictions. The model takes a total of four features and two different target variables. Once inputs have been completed, running the model would provide a real-time output (Figure 2) of the expected customer response or engagement rate based on the selected survey incentives and survey completion time, for a particular industry and survey. Additionally, the model will suggest alternate survey incentives and survey completion time (optional) which are predicted to get a better reaction to the dependent variable of the user's choice.

We provide two sets of three recommendations based on the users' needs. In the first set of recommendations, the model keeps the survey length fixed, and provides recommendations only on the survey incentives. In the second set of recommendations, the model provides recommendations on both alternative survey incentives and alternative survey length. For campaign engineers, our model provides them a very flexible solution based on their needs.

VI. ALGORITHM USED

This algorithm used in the survey incentive model is a Random Forest Regressor. In this model, the random forest regressor creates a set of decision trees from a randomly selected subset of training set and then it collects the votes from different decision trees to decide the final prediction.

To improve the predictive accuracy and control over-fitting, the bootstrap and the sub-sample size are also ensembled as two important parameters.

First, Bootstrap sampling is used when building models to repeatedly sample data with replacement from the original training set to reduce the variance of the predictions, thus greatly improving the predictive performance. The model randomly selects a fixed percentage of the whole training set with replacement as a bootstrap sample and grows a decision tree from the bootstrap sample.

```
Dependent Variable: Conversion_Rate
Industry Type: Retail
Survey Type: Customer_Survey

Your Current Conversion_Rate is 2.4250000000000003 % with $11-$20 incentive amount and a 0-5 minutes survey.

If want to keep the current survey length, then the model recommends:
$ 34.0 , Conversion_Rate is 2.8000000000000003 %
$ 35.0 , Conversion_Rate is 2.6 %
$ 32.0 , Conversion_Rate is 2.4 %


If you are open to a different survey length, then the model recommends:
$ 30.0 , 8 minute(s), Conversion_Rate is 5.4 %
$ 41.0 , 13 minute(s), Conversion_Rate is 3.292 %
$ 52.0 , 30 minute(s), Conversion_Rate is 3.0 %

Model accuracy:  98.87204613892717 %
```

Figure 2. Model output overview

To improve the predictive accuracy and control over-fitting, the bootstrap and the sub-sample size are also ensembled as two important parameters. First, Bootstrap sampling is used when building models to repeatedly sample data with replacement from the original training set to reduce the variance of the predictions, thus greatly improving the predictive performance. The model randomly selects a fixed percentage of the whole training set with replacement as a bootstrap sample and grows a decision tree from the bootstrap sample. Second, feature subsampling randomly selects subsets of features considered when splitting nodes in each decision tree. At each node, the model randomly selects "d" features without replacement and then splits the node using the feature that provides the best split according to the objective function, for instance, by maximizing the information gain. In the end, the model aggregates the prediction by each tree to assign the class label by majority vote. The implementation of feature subsampling prevents model overfitting effectively.

VII. MODEL VALIDATION

To evaluate the performance of the prediction model, the regression model is trained with a subset of the dataset while it is validated with the remaining. The model is trained using 80% of the dataset while the performance is evaluated using the remaining 20%. This partitioning ensures more samples for training while giving a sufficient number of samples for validation. An increased number of training samples can potentially reduce any over-fitting in the model.

We apply different regression models, and we use the R-square score as a validation score to evaluate the model's performance. R-square (R^2) score, also known as the coefficient of determination, is a statistical measure of how well the linear model explains the variable variation. In our case, the r-square score provides insights into the model performance on the data. The R-square (R^2) score ranges from 0 to 1, and the higher the score, the better the model performance is.

The result from the model tests shows that the random forest regressor performs the best in predicting the character counts and the engagement rates, with a 0.9887 r-square (R^2) score. The Evaluation scores are displayed in Table 2.

Model	R-square Score
Linear Regression	0.398
Polynomial (degree=3)	0.52
XGboost	0.92
DecisionTreeRegressor	0.9887

Table 2. Model Performance

VIII. ASSUMPTIONS

There are several assumptions that had to be made during the entire process. For the survey incentives, the rewards values must not be $0 or any other form. Because the model takes the value of the rewards as an input, any other forms of survey incentives need to be converted to the incentives dollars to get the proper recommendations. We also assume that the campaign engineers the survey length has an impact on the email click-to-open rate and conversion rate. Based on these 2 assumptions, the model takes the survey incentives and survey lengths as 2 independent features to help campaign engineers manage email engagement.

IX. USE CASES

The survey incentive model is developed for email marketing campaigns, particularly for survey email campaigns. It is to be used by the campaign engineers within the workflow of the campaign and to identify ways to increase user engagement prior to deployment based on given parameters or inputs. The current model provides predictive analytics for click-to-open rate and conversion rate.

The end-users are the email campaign marketing teams whose goal is to increase email engagement rates, such as click-to-open rate, and conversion rate, etc.. These target variables can be very different based on the campaign engineer's choice and can be customized based on a variety of different inputs.   Our example in this report takes into consideration only a few inputs. Once the campaign engineers decide on the type of survey included in an email campaign, the model will calculate and serve the predicted user response to the survey incentives and survey length in real-time.

The interactive UI shown in Figure 3 (demo purposes only) allows campaign engineers to try different options regarding survey incentives as well as survey length. This allows the campaign engineers to quickly and accurately know if users will find the survey incentives engaging or not, before sending the email out. Since the predictions are served in real-time, the campaign engineers will be able to run many survey incentives and survey length scenarios for modeling purposes. Combining more than one   model in the Loxz family of email recommendations to tune their emails to their audience without delay, and improving engagement rates is also another option.

Figure 3. Model interactive settings

X. DATA SCIENTIST WON CONCLUSION

As the model was trained on real-world campaigns and survey data, the effectiveness of our predictions and recommendations will be directly correlated to the email engagement rates of alternative survey incentives and survey lengths.

The model is constructed to predict the customer response to the survey incentive and length of a survey email campaign, for the purpose of improving campaign email engagement rates. We showed the model to have a high accuracy (98.87%), which indicates the model has the ability to make useful and timely predictions for the users. We also showed the algorithm serves the prediction in less than half a second, which makes this project eligible for use in real-time applications. The ability to provide an accuracy score to users enables campaign engineers to trust the usage of the model. And it also provides a way for campaign engineers to validate their offers.

ⅩⅠ. FEATURE CONSIDERATIONS

For immediate future work, the model will be extended to predict Revenue-per-email values. However, each campaign engineer might want to optimize for a particular feature in the dataset and can be materially effective, for conversion rates as well. Furthermore, different forms of the survey will be included in the training process. However, including additional survey incentive forms, the model will need to be further developed to filter out the irrelevant audiences of different engagement metrics. According to a study by Getsitecontrol, "survey incentives can also attract the wrong crowd – meaning, those who might not be your target audience"[7]. For example, in our case, offering amazon gift cards will probably attract more online shoppers, but if the target audiences of your email campaign are from academic and educational institutions, the current model cannot distinguish those relational factors. In the future version of the model, we are considering introducing the cross relations between different forms of survey rewards and the survey type and industry.

It is also important that if you combine two or more models from the Loxz Portfolio, the possibility of enhanced engagement rates are prevalent. For example, if you including a Sentiment Analysis Model or a Character Count model and run those models simultaneously, in a multi-modal environment, you may find increased engagement rates.

REFERENCES

[1] Prince, H. (2021, December 8). What is the best survey reward? let your respondents decide. Rybbon. Retrieved August 29, 2022, from https://www.rybbon.net/blog/what-is-the-best-survey-reward-let-respondents-decide/

[2] Holmes, K. (2022, July 13). Average conversion rate by industry and marketing source. Ruler Analytics. Retrieved August 30, 2022, from https://www.ruleranalytics.com/blog/insight/conversion-rate-by-industry/

[3] Ameet V Joshi. 2020. Amazon's Machine Learning Toolkit: Sagemaker. In *Machine Learning and Artificial Intelligence*. Springer Nature, Chapter 24, 233–243. https://doi.org/10.1007/978-3-030-26622-6_24

[4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830. http://scikit-learn.sourceforge.net.

[5] Jupyter widgets. Jupyter Widgets - Jupyter Widgets 8.0.1 documentation. (n.d.). Retrieved August 30, 2022, from https://ipywidgets.readthedocs.io/en/stable/

[6] Boto3 documentation¶. Boto3 documentation - Boto3 Docs 1.24.62 documentation. (n.d.). Retrieved August 30, 2022, from https://boto3.amazonaws.com/v1/documentation/api/latest/index.html

[7] Boutin, C. (2021, March 18). How to use survey incentives to increase response rates. Getsitecontrol. Retrieved August 30, 2022, from https://getsitecontrol.com/blog/how-to-use-survey-incentives/