Prediction of Bank Marketing Campaign Success

Claire Ellison-Chen (te2049@nyu.edu), Siyu Shen (ss14359@nyu.edu), Zhuoyuan Xu (zx1137@nyu.edu)
Center for Data Science, New York University

Business Understanding

Marketing campaigns are a substantial method to enhance business operations of banking institutions by directly reaching out to potential clients and building up two-way, individualized and long-term connections with them [1]. Such connections have become more important than ever in this fast-growing generation. Whilst the development of the industry has been continuously attracting new players [2], intensifying peer competition in a limited space, social progress drives more individuals and corporate customers to seek banking services. Facing these challenges and opportunities, banks have been raising marketing expenses steadily in recent years, and the top 25 leading U.S. banks grew marketing spending by 7% in 2019 to $15.4 billion [3]. It is pressing for banks to learn how to optimize their marketing costs on clients who would yield more revenue.

With the enormous amount of client data collected during regular business, supervised learning models enable banks to predict outcomes of marketing campaigns before they take place, hence curtailing the operational costs of unproductive contacts and uncovering the characteristics of clients who are likely to subscribe to their services. Our project is to build a predictive tool for the telemarketing campaign of a primary bank service "term deposit". Classification and prediction-based data mining solutions on similar problems of other marketing strategies such as emails [4] have been widely explored in previous studies and put to production via designated softwares like cloud analytics [5]. We plan to compare the performances of various classification models, and the one with the best predictive power to identify successful subscribers would be deployed into production to recognize profitable contacts in new marketing campaigns.

Data Understanding and Preparation

Data Source and Variable Description

The bank marketing dataset in this study is from the UCI Machine Learning Repository [6] and originated from a telemarketing campaign, during which responses were collected from direct phone calls to the existing and potential clients of a Portuguese institution. Each bank client was offered a subscription opportunity to the term deposit during the campaign. More than one contact with the same client was required in order to assess if the term deposit is successfully opened.

Independent Features

The dataset contains instances of 41188 bank clients and 20 features in total, ordered by date (from May 2008 to November 2010). Features consist of both external information and the internal data from the bank. The internal features consist of client demographic data and campaign-specific records. Each of these features carries unique characteristics of the client and thus all except ones risking data leakage are kept for the modeling process, as the project prioritizes the predictive power of the model. The external data includes social and economic context attributes that are indicators of the overall economy. To differentiate the effects of internal and external features, we use internal features to build up our baseline models, as we specify in the modeling section.

Target Variable

The target variable in our prediction is whether the client makes a term deposit. We encode 1 for 'yes' and 0 for 'no' in this categorical binary variable. Aligning with the common expectation of telemarketing, the successful subscription rate turns out low and only takes 11.3% of all the records. Class imbalance carries intrinsic threats to our classification model. Due to the overpopulation of the negative cases, the trained model would be less likely to predict positive. We have to keep this issue in mind while designing and evaluating our algorithms.

Bias in Telephone Samples

Telephone surveys are known to face difficulties in the underlying sampling bias that may lead to unbalanced data or require corrections, careful causal inferences and limited generalization depending on the actual problems to solve. Firstly, each person faces considerably more exposure to promotions as marketing campaigns become more popular. Secondly, people are wary about their information security on unknown calls. As a result, some people tend to ignore promotions or provide negative and fake feedback without listening to call details to avoid future survey requests, even though they may need the service. These might create a large amount of rejection or unknown data points in the dataset as well as non-response bias. Meanwhile, the dataset can be less able to thoroughly reveal the characteristics of people who may subscribe. In consideration of such biases, our group takes careful steps to generalize our conclusion, deal with missing values, and balance our dataset by upsampling.

Data Cleaning

The first step is to check missing values in the dataset. There are no null values, and all missing values are indicated by a categorical value: "unknown". Many features contain "unknown" values of non-negligible portion, especially the variable about credit in default which has 20.9% "unknown" value. Still, "unknown" carries useful information as some clients might deliberately withhold their response to certain questions. Hence, we decide to not replace "unknown" with other values and keep it as a dummy variable.

The *Duration* feature represents the last contact duration and was not known before making a call. Although it is highly correlated with the target variable, it risks potential data leakage, and only serves as a benchmark for checking rather than predictive factor in our study. Thus, we decide to drop it.

After dropping *Duration*, we make dummy variables of all categorical features except for *pdays*. This numeric feature represents the number of days since the last contact of a previous campaign, and has a special outliner of 999 for no previous contact. We transform this feature to categorical dummy variables by binning the number of days into categories of every 5 days, and 999 for its own category. The rest of the numeric features are either number counts which are all less than 100, or economic and social index values which are small and with low variance. We standardize them for logistic regressions which require features to be normalized, but keep them as they are for decision trees and ensemble models which are not sensitive to variance in data and do not require scaling.

Modeling & Evaluation

Evaluation Metrics

Due to class imbalance, classification accuracy can be misleading, as our models are likely to predict most cases in the majority "no" class correctly. However, we are more interested in classifying clients who would make a term deposit. Here, false positive and false negative errors can both be problematic in business practice. We want to avoid false negative cases, in which the clients who actually subscribed to term deposits are labeled as "no", to prevent losing potential deposits. If the model predicts too many false positive cases, the marketing campaign will waste budget and time on clients who are labeled as "yes" but do not make term deposits. Therefore, we plan to take the following evaluation metrics to form a comprehensive evaluation process:

a. Area Under the ROC Curve (AUC)

Instead of using accuracy, we choose AUC score to measure the overall predictiveness of classifiers over all possible thresholds.

b. Recall

Recall measures the true positive rate. The business case demands us to maximize recall as so to identify as many clients who would make a term deposit as possible.

c. Precision

While maintaining a high recall score, we want to avoid low precision scores. The inflation of false positives would lower precision, giving an alarming signal that many predication results as "yes" are valueless to the marketing campaign.

d. F1-score

As the weighted average of precision and recall for a specific probability threshold, F1-score can provide additional information about the balance and accuracy in the recall-precision trade-off.

Overall, we will first rate each classifier by AUC score, then focus on its ability to predict the positive class by prioritizing recall, complemented by precision and F1-score for the final model selection.

<u>Algorithm Choices</u>

a. Baseline: Logistic Regression Model

We select the Logistic Regression model as the baseline for this classical binary classification problem. It is a robust and simple model generally resistant to overfitting, which makes it ideal for a baseline.

We select four other models to compare with the baseline model:

b. Decision Tree

In our study, the classifier would benefit the marketing department in production because it is inexpensive to construct and fast for inference. It also automatically excludes unimportant features for feature selection and easy to interpret. However, it risks overfitting, and small changes in training set can lead to large changes in trees, which might raise concern and variations when the department has to periodically retrain the model with new training data. Although it does not require much data preparation, it does not support missing values which are likely to appear in our dataset in the real world practice. Therefore, missing values must be taken care of during data preparation [7].

c. Random Forest

As a bagging ensemble algorithm, Random Forest considers various decision trees to form the final result. Hence, it omits Decision Tree's overfitting issue, and would lower overall variance and errors of the trees. It provides high accuracy on imbalanced dataset, like our case here. It can intake large amounts of data in industry practice and handle missing data better than Decision Tree [8]. However, the algorithm loses the tree's interpretability and appears to be a black box method. It suits a predictive project better than a descriptive one. Also, long training time needs to be reserved ahead in production.

d. Gradient Boosting

The boosting ensemble algorithm builds trees sequentially with the later tree correcting errors of trained trees. Although it may take longer training time, it could yield better results than Random Forest and curbs overfitting. However, it requires careful tuning which might be a time efficiency concern in production. It is also hard to interpret and small changes in the training set can radically change the model [9].

e. Ada Boosting:

Compared to Gradient Boosting, this boosting ensemble algorithm is faster and easier to use with few parameters for tuning. Its sensitivity to noise and outliers is a disadvantage in general but might benefit in our case to identify the minority class better [10].

Improvement Upon Baseline

Step 1. Performance with Default Setting

We start with the first hypothesis that using the internal bank data of clients could predict their subscription to term deposit, as opening a term deposit at the given bank is a personal financial decision associated with demographic information, banking histories, and campaign outreach. After an 80-20 train-test split on our dataset of corresponding features, we performed all default classifiers on the training set and then recorded evaluation metrics from the test set.

	AUC Scores	Recall	Precision	F1
Logistic*	0.754	0.17	0.62	0.26
Decision Tree	0.604	0.29	0.28	0.29
Random Forest	0.735	0.22	0.49	0.31
Gradient Boost	0.766	0.19	0.62	0.29
AdaBoost	0.759	0.17	0.61	0.27

*Baseline Model

Table 1. Evaluation Results of Default Models

With most AUC scores beyond 0.7, our models have good predictiveness to start with. The low recall scores are probably caused by class imbalance.

Step 2. Imbalanced V.S. Balanced Dataset

Class imbalance hinders our classifiers' ability to predict the positive class. If a model is trained on a balanced dataset, it will learn the better decision boundary between the 2 classes and improve its performance. To rebalance the dataset, we upsample the minority class of "yes", instead of downsampling which would cause us to lose too much data (around 80%) that could have provided extra information or avoided outliers.

We introduced the Synthetic Minority Oversampling Technique (SMOTE) to upsample and rebalance our dataset. The method synthesizes new samples from the minority class and causes the classifier to build larger decision regions that contain nearby minority class points, in opposition to simple upsampling which can balance the class distribution but does not provide any additional information to the model [11].

After the 80-20 train-test-split, we performed SMOTE on the training set to achieve class balance, then trained all default classifiers on the new upsampled set. The evaluation metrics from the test set are as below:

	AUC Scores	Recall	Precision	F1
Logistic*	0.754	0.18	0.61	0.28
Decision Tree	0.604	0.30	0.27	0.28
Random Forest	0.748	0.31	0.44	0.36
Gradient Boost	0.761	0.43	0.45	0.44
AdaBoost	0.729	0.36	0.45	0.40

Table 2. Evaluation Results after Data Balancing

We witness significant increase in recall scores in all ensemble models, although these of the Logistic regression and Decision Tree do not fluctuate much.

The AUC scores remain similar among all classifiers with a small margin of increase or decrease. We decide to keep the upsampling of training set as the first step of the pipeline in the following practice.

Step 3. Additional Features

Adding features might further improve accuracy of the models. So far we only used internal data from the bank to predict the client's term deposit subscription. Our second hypothesis is that the economic and social index can contribute to our prediction, since the social and economic environment might impact personal behaviors. After the train-test split, SMOTE, and model training, the new results are:

	AUC Scores	Recall	Precision	F1
Logistic*	0.785	0.22	0.60	0.32
Decision Tree	0.627	0.35	0.30	0.32
Random Forest	0.771	0.35	0.49	0.41
Gradient Boost	0.776	0.43	0.49	0.45
AdaBoost	0.758	0.45	0.44	0.45

Table 3. Evaluation Results after Adding Social and Economic Features

We witness an increase in AUC scores and recall over all classifiers. New features of social and economic index have provided additional information to our model.

Step 4. Hyperparameter Tuning with Validation

We use grid search for hyperparameter tuning with 5-fold stratified cross-validation which makes each fold an appropriate representative of the original data concerning the class imbalance. The scoring metric is recall, as we hope our final model to have strong predictive power for the positive class.

Important hyperparameters tuned for each model:

1. **Logistic regression:** *penalty = l2* for regularization method and *C=0.001* for penalty strength.
2. **Decision Tree:** *min_samples_split = 2000* and *min_samples_leaf = 20*: complement each other and control the depth of the model and complexity.
3. **Random Forest:** *n_estimators = 100:* controls the model computational complexity; *max_features = log2*: defines how many features each tree is randomly assigned and controls overfitting; *min_samples_leaf = 150*: controls the growth of the trees and prevents overfitting.

4. **Gradient Boosting:** Macro-level hyperparameters *n_estimators=20:* the number of trees and *learning_rate = 0.05:* shrinks the contribution of each tree. Together they control the learning ability and computational complexity, and need to be tuned to prevent overfitting. Tree-level parameters *max_depth, min_samples_split* are tuned but not very helpful in our case.
5. **Ada Boosting:** *n_estimators = 0.05* and *learning_rate = 25*: similar with gradient boosting. More trees may require a smaller learning rate and vice versa.

Best performance after tuning:

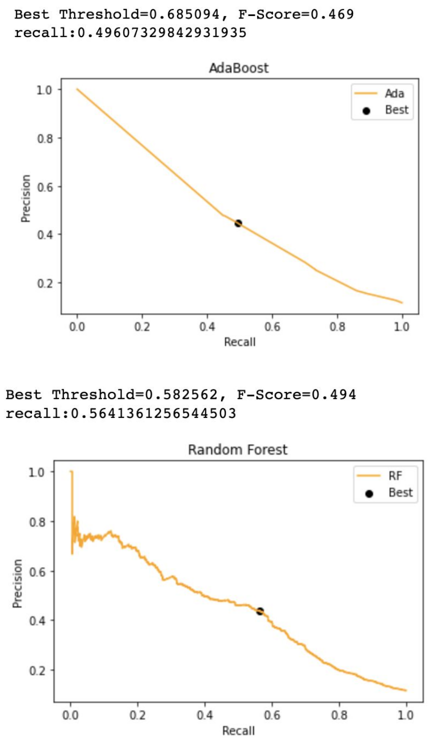
	AUC Scores	Recall	Precision	F1
Logistic*	0.786	0.33	0.55	0.414
Decision Tree	0.770	0.547	0.404	0.463
Random Forest	0.788	0.596	0.410	0.486
Gradient Boost	0.782	0.583	0.402	0.476
AdaBoost	0.771	0.692	0.269	0.387

Table 4. Evaluation Results after Tuning

All the other models outperform the baseline in recall while sharing similar slightly-increased AUC scores. AdaBoost gives the best recall, but overall Random Forest is our best performing algorithm.

Step 5. Recall-Precision Threshold

We want to check if our results still have room to improve. Our model is optimized for recall in the precision and recall trade-off, while F1-score reflects the balance between them. We hope to see if we can increase our recall reasonably while maintaining a satisfactory F1-score. Hence, we marked the optimal point for F1-score on the precision-recall curves of our final candidate classifiers AdaBoot and Random Forest.



```
Best Threshold=0.685094, F-Score=0.469
recall:0.49607329842931935
```

```
Best Threshold=0.582562, F-Score=0.494
recall:0.5641361256544503
```

Figure 1. Precision against Recall. Above: Ada Boost; Below: Random Forest

In Random Forest, our current F1-score is already close to the optimal value, yet would sacrifice recall if we shift the threshold to match it. If we change the threshold significantly in AdaBoost, the weak F1-score can be improved but the model will lose its advantage of high recall. Therefore, changing the threshold in both models will not provide additional benefit to our model. We will keep the default 0.5 threshold and have the models as final.

Model Selection

We have obtained two models Random Forest and AdaBoost to help the campaign predict clients who would make a term deposit. The telemarketing team can input a list of prospective contacts and receive predictions on whether the model believes they would make a term deposit by the end of campaign. The team can decide to remove clients from the contact list who are unlikely to subscribe and use their time and budget more responsibly on clients who would draw profits to the bank.

Random Forest is a well-balanced model to show good performance on the AUC scores, recall and precision. As in the confusion matrix of the test results from our final models, it helps filter out most negative cases. However, although AdaBoost has weaker predictiveness than Random Forest in general, it specializes in capturing more potential subscribers, albeit at the cost of having much more false negative predictions.

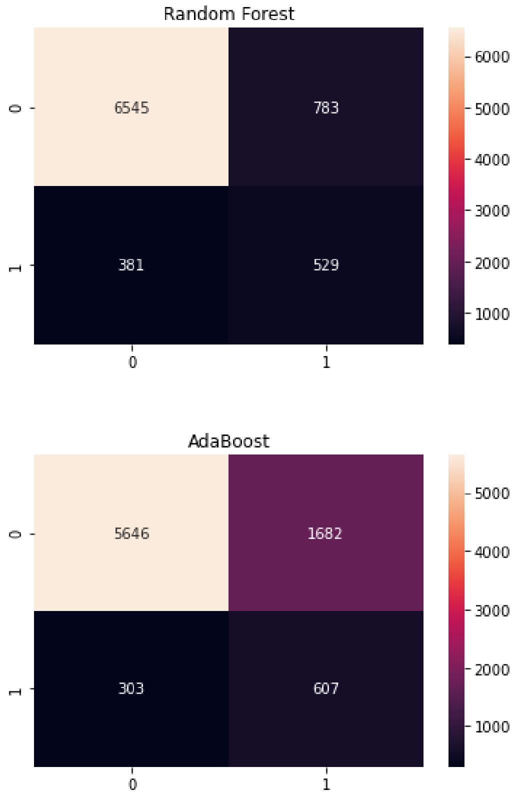


Figure 2. Confusion Matrix Heat Map. Above: Random Forest; Below: AdaBoost

The two models suit different priorities of the telemarketing campaign. If the bank wants to seize every possible term deposit, AdaBoost will help it identify more clients who would subscribe. However, the campaign would also reach out to 977 more clients than if using Random Forest. Due to constraints of labor and call hours, Random Forest will be a more efficient choice if under a limited marketing budget.

Deployment

Although we used a prepared dataset in this project, in an agile environment of business practice, we need to create an automated system that can retrain the model periodically and keep up with updated information from completed campaigns. The data mining system and the model shall be deployed together to meet ad hoc needs:

1) **Data Integration:** Economic index data needs to be collected periodically and stored in the institution's data management system along with the internal bank data. Existing client's features will be updated after every marketing campaign. When a new campaign begins, the marketing department would pull the list of existing and new clients as prospective contacts by id or timeframe for the next steps.
2) **Data Cleaning Pipeline**: The raw data is then transformed into structured features in the data cleaning pipeline ready for modeling.
3) **Train and Validation**: Run the Train/Validation module on the cleaned and engineered data. This is an automated system for testing and periodically updating the prediction model. It then creates and stores the new prediction model.
4) **Model:** This holds the actual prediction models (Random Forest or AdaBoost)
5) **Prediction:** The institution uses the new prediction model to classify target clients for the current marketing campaign.

Business Values

The institution can benefit from the predictive model to increase profits and enhance operations in the following ways:

1) **Clientele Selection:** the model prediction would return a target client group who are likely to subscribe to a term deposit for telemarketing. This would possibly increase client response rate and client satisfaction if they're offered the desired product.

2) **Budget Control and Labor Productivity**: the institution can directly reach out to this target group and reduce costs of redundant marketing outreach as well as improve its operation efficiency.

3) **Campaign Extension:** the institution could include the target clients in future marketing campaigns which markets follow-up services beyond term deposit.

From feature importance results, social and economic features contribute the most to the model performance (see Appendix A). Some top deterministic features from the internal bank data are: *contact_telephone*, *default_unknown*, *poutcome*, *month_may*, *job_blue-collar*. Based on the results and feature correlations, some recommendations can be made to the institution:

- The institution should be aware of the economic trend and tune the campaign accordingly to the trend. Clients tend to be more likely to subscribe when the overall economy improves (i.e. when Consumer Confidence Index and Consumer Price index increase).
- There is a potential seasonal pattern of client's willingness to open term deposits for planning campaign schedules.
- As the method of contact is a strong indicator of clients' decision, investigating telephone and cellular users' information, possibly through unsupervised learning or more descriptive modeling, might provide valuable insights.
- The institution should focus more on the clients who have previously subscribed to a term deposit.

Ethical Considerations

The bank needs to be mindful of the clients' right to privacy. Our model relies heavily on customer's demographic information and banking histories to make predictions. If the bank acquires parts of such information from a third party, or when the prediction results direct the telemarketing campaign to contact new clients, some targeted prospective customers would haven't requested information about the product. They may not want unsolicited marketing and even consider it unethical [12].

Risks

Some costs may occur in the institution: 1). the expense of calling and marketing, 2). Annoyed clients who are not willing to open term deposits. To mitigate these costs, we should retrain the model periodically , test new features, and compare our predictions with the actual results after each campaign, perhaps using AB testing to compare groups' subscription success with and without direction from the model prediction results.

The result of our model only classifies a target group who are likely to subscribe through the telemarketing campaign, but we can't consider those clients as our entire target group for term deposit. Other clients may participate in other kinds of marketing such as digital marketing or on-site promotion. In order to target a more comprehensive group of customers, the bank may consider extending the model to generalize the prediction of term deposit subscription by including new features from other marketing methods.

References

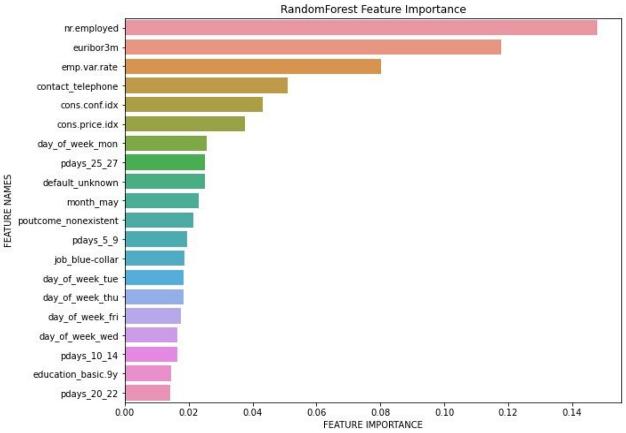
1. Rust RT, Moorman C, Bhalla G (2010) Rethinking Marketing. In: Harvard Business Review. https://hbr.org/2010/01/rethinking-marketing. Accessed 30 Nov 2020
2. Nurun (2013) The 21st Century Bank: Rethinking and Transforming Financial Institutions. In: Nurun Perspective. https://www.nurun.com/en/global/perspective/reinventing-the-bank/the-21st-century-bank-rethinking-and-transforming-financial-institutions. Accessed 30 Nov 2020
3. Emiboston.com. 2020. Steady Growth In Marketing Spend And Marketing Ratios For Top U.S. Banks In 2019 | EMI Strategic Marketing. [online] Available at: <http://emiboston.com/top-u-s-banks-grew-marketing-spend-and-marketing-ratios-in-2019/>
4. Abakouy R, Mokhtar EE, Haddadi, AE (2017) Classification and Prediction Based Data Mining Algorithms to Predict Email Marketing Campaigns. The 2nd International Conference
5. Kuttikat, J (2017) Add Machine Learning For an Effective Marketing Campaign. In: Oracle Analytics Advantage. https://blogs.oracle.com/analytics/add-machine-learning-for-an-effective-marketing-campaign. Accessed 30 Nov 2020
6. S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
7. Medium. 2020. Decision Tree Classification. [online] Available at: <https://medium.com/swlh/decision-tree-classification-de64fc4d5aac> [Accessed 4 December 2020].
8. Gupta, S., Feb 2020. Pros And Cons Of Various Classification ML Algorithms. [online] Medium. Available at: <https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6> [Accessed 4 December 2020].
9. Yildirim, S., 2020. Gradient Boosted Decision Trees-Explained. [online] Medium. Available at: <https://towardsdatascience.com/gradient-boosted

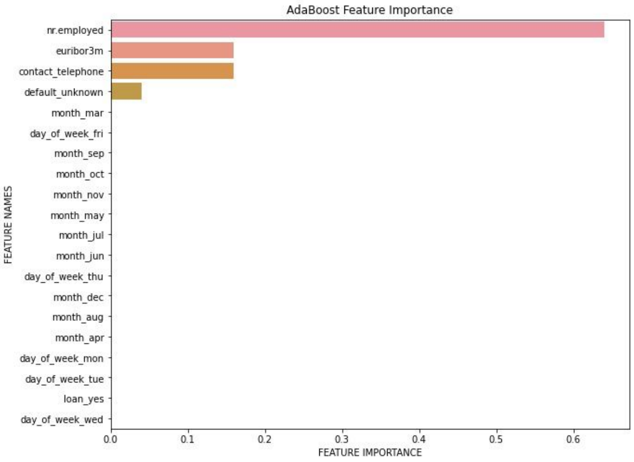
-decision-trees-explained-9259bd8205af> [Accessed 4 December 2020].

10. Kurama, V., 2020. A Guide To Understanding Adaboost | Paperspace Blog. [online] Paperspace Blog. Available at: <https://blog.paperspace.com/adaboost-optimizer/> [Accessed 4 December 2020].
11. Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, pp.321-357.
12. Savannah, Samoszuk (2013) Direct Marketing: Legal and Ethical Issues https://study.com/academy/lesson/direct-marketing-legal-ethical-issues.html#:~:text=Ethical%20issues%20include%20violating%20the,laws%20that%20effect%20direct%20marketing.

Appendices

Appendix A. Feature Importance of Two Final Models





Contributions

- Zhuoyuan Xu: Business Understanding, Data Understanding, Data Cleaning, Model Design
- Claire Ellison-Chen: Evaluation Metrics, Improvement upon baseline
- Siyu Shen: Model Design, Model Evaluation, Deployment, Ethical Considerations and Risks