Social Media Ads Marketing Strategy Optimization with Optimal Regression Tree and XGBoost

Project Final Report
Xinyao Han and Claire Guan

Introduction

Social media advertising is a type of digital marketing in which paid ads are delivered to the target audience via social networks such as Facebook and Instagram, which provides a quick and effective way to connect with customers and boost marketing campaigns. And one challenge in ad campaigns is to minimize the conversion cost associated with the advertisement, that is, the total cost of conversions by the total number of conversions.

In recent years, models from logistic regression to Deep Factorization Machine have been used to address this critical business problem. To minimize the total ads spend for various campaigns, this project have two aims: (1) predict the cost per conversion and analyze the user groups with the lowest cost per conversion, (2) understand the relationship between conversion cost and user demographics, and other interaction behaviors.

Dataset

Our dataset is a Kaggle *Sales Conversion Optimization* (see Appendix) dataset, which contains 1143 entries from a Facebook advertisement campaign of Company XYZ (an anonymous organization). It focuses on three main campaigns of Company XYZ, Campaign 916, Campaign 936, and Campaign 1178. The data is aggregated in consideration of user privacy. Each entry is an aggregated user group with group characteristics of age, gender, and interest. We have 11 features which include advertisement features, user features, and user response features.

(1) Advertisement
Ad ID categorizes each ad, *Facebook Campaign ID* categorizes each ID using FaceBook's standards, and *Ad Spent* is the cost the company paid to show an Ad.
(2) User
We use *age, gender*, and *interest category* to define a user group. A user group is an aggregated group of users with the same age range, gender and interest category. The feature *interest* is a code specifying a user's interest, which is inferred from one's Facebook profile.
(3) User Response
Ad impressions, clicks, total conversions, and *approved conversions* are features that we define as different stages of an ad promotion. *Ads Impressions* are the number of times an ad was shown, *Clicks* are the number of users who click on an ad. *Total conversion* is the number of inquiries about the product. *Approved conversion* is the total number of people who bought the product after seeing the ad.

We generate more features based on the logic of advertising placement, which is analyzed in the exploratory data analysis section. All the features are used for the prediction task.

Exploratory Data Analysis

We conducted some exploratory data analysis to understand the user demographics (age, gender, interest) and cost of conversion in our dataset. In this project, we consider the conversion to be a purchase, and we defined three metrics - ClickRate (Clicks/Impressions), Conversion Rate (Purchase/Impressions), and cost per conversion (Cost per conversion - Spent/Purchase) to measure campaign marketing effects. We selected two campaigns (No.936 and No.1178) with the most data, which targets different age, gender, and interest user groups. Campaign No.936 has 464 identified user groups, and campaign No.1178 has 625 user groups.

(a) User Group Demographics

1. Age distribution**.** Overall 37% of user groups are between 30-34, 21.7% of user groups are between 35-39, 18.4% of user groups are between 40-44, 22.7% of user groups are between 45-49.
2. Gender distribution. Overall we have a balanced gender distribution of 51.2% male and 48.8% female.
3. Interest distribution. Overall, there are 40 identified interest groups, among them, Interest 16 takes up 12.3%, Interest 10 takes up 7.4% and Interest 29 takes up 6.74% and the others contribute between 0%-5%.

(b) Conversion Rates

1. Campaign No.936 has an average conversion rate of 0.005%, Campaign No.1178 has an average conversion rate of 0.001%
2. Campaign No.936: higher conversion for male users between 35-44
3. Campaign No.1178: higher conversion for male and female users between 30-34

(c) Conversion Cost

1. Campaign No.936: Average conversion cost 8.44, ranging from 2 to 24 per user group
2. Campaign No.1178: average conversion cost 58.86, ranging from 31 to 116 per group

(d) Correlation Heatmap

In addition, we plot the correlation heatmap (before converting to dummy variables and standardizing) below.

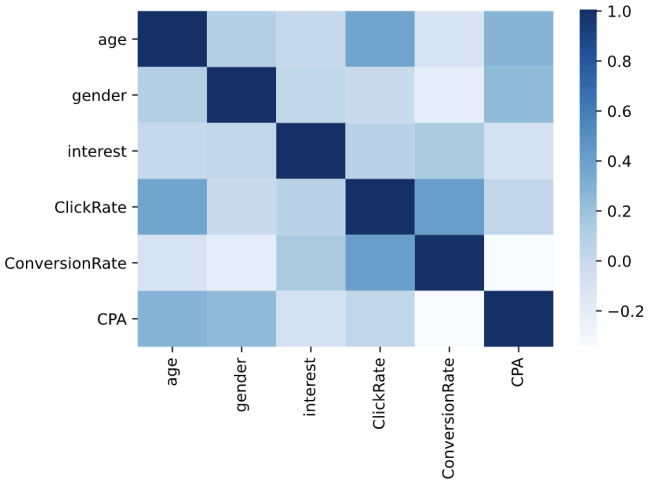


Figure 1. Feature Correlation Heat Map

We can see that ClickRate is positively related to Conversion Rate, ConversionRate is negatively correlated with cost per conversion (Cost of conversion), and cost per conversion is positively correlated with age, gender, and interest, which aligns with our hypothesis.

From the EDA, we conclude that these campaigns target different user groups, therefore we build models for the two campaigns No.936 and No.1178 separately.

Methodology

Our task is to optimize cost per conversion to get minimized cost per conversion user group. We divide this task into two stages.

- In the first stage, we capture user group purchasing behavior by predicting the cost per conversion, where we implemented four models: *Lasso Regression, Holistic Regression, ORT-L*, and *XGBoost* to predict the conversion cost for different user groups based on user characteristics and interactions for two campaigns respectively.
- In the second stage, we utilize Interpretable K-Means Clustering to identify the key features influencing the conversion cost and to segment customers in order to derive the target groups for different campaigns. We ran an OCT after K-Means Clustering for interpretability.

1. Prediction on cost per conversion
We standardize the training data with MIN MAX scaler. Min-max scaler subtracts the minimum value of the feature from the feature and divides it by the range. It preserves the shape of the feature and could also deal with categorical features that use numbers to represent.

(a) Lasso Regression.
We run the Lasso Regression model to predict the cost per conversion as a baseline model. Lasso regression is a type of penalized regression method. It's used for a large number of features because it selects a subset of import features by shrinking some parameter estimates to zero. It provides greater prediction accuracy compared with linear regression because it fights against overfitting. Lasso regression uses the L1 norm as a penalty term. It is represented using the following formula:

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1$$

Using the standardized training data, we tune the hyperparameter λ with cross-validation. We use the best λ = 0.2 to train the model. The results will be discussed in the next section.

(b) Holistic Regression.
We run the Holistic Regression model to predict the cost per conversion and compare the performance with the other models. We add interaction terms between user group characteristics and user response behaviors on the original dataset. We create interactions by respectively multiplying *age, and gender* with *impressions, clicks,* and *total conversions*. We also normalize the data using Min Max scaler after creating the interaction terms. We transform the data by replacing each feature with the following transformation: squared feature, square root of absolute feature, and take algorithm of feature value (ϵ=1):

$$\mathcal{T}_j \;=\; \{\tilde{X}_{4(j-1)+1} := X_j,\ \tilde{X}_{4(j-1)+2} := X_j^2,\ \tilde{X}_{4(j-1)+3} := \sqrt{|X_j|},\ \tilde{X}_{4(j-1)+4} := \log\left(|X_j|+\epsilon\right)\},$$

We derive a matrix of features with pairwise correlation using Pearson correlation. We add constraints to model robustness, sparsity, nonlinear transformations, and pairwise colinearity. We model robustness with the regularization term λ||β||. We model sparsity by requiring at most k features of all the variables selected by the model. We model nonlinear transformations by selecting at most one transformation of each feature is selected. We model pairwise colinearity by selecting at most one feature from feature pairs with high pairwise correlation. The model formulation is attached in the appendix. We tune the model by using a grid-search on k, λ, the correlation parameter ρ. We achieve the best result with k = 4, λ =0.2, and ρ = 0.9. We will discuss the results in detail in the next section.

(c) ORT-L.
We run the Optimal Regression Tree (ORT-L) with linear predictions to predict the cost per conversion, which is an ORT implementation used here for better interpretability and data adaptability. ORT-L is an example of the Optimal Trees methodology applied to regression problems, where the prediction in each leaf is for a continuous outcome. In addition, ORT-L is constructed in a way that each leaf uses a linear model to make its predictions. To avoid overfitting, we penalize the norm of the regression coefficient vector in each leaf, by including a regularization parameter λ.

$$\frac{1}{\hat{L}}\sum_{i=1}^{n} L_i + \alpha \cdot C + \lambda \sum_{t\in\mathcal{T}_L}\sum_{j=1}^{p} r_{jt}.$$

The model formulation is attached in the appendix. We first select the tree depth=4 using a simple ORT using grid search. We then tune the model by looping over λ values of 0.005, 0.01, 0.05,0.1. We achieve the best result with λ=0.05.

(d) XGBoost.
We run the XGBoost model to predict the cost per conversion and compare the performance with the other models. XGBoost (eXtreme Gradient Boosting) is a popular ensemble boosting tree method that has features such as the clever penalization of trees and proportional shrinking of leaf nodes. We use the sklearn implementation of the XGBoost model, and fine-tuned its parameters. We achieve the best result with max_depth=4, n_estimators=100, and learning_rate=0.01.

2. Customer Segmentation with Interpretable K-Means Clustering
To better understand the factors that are critical for different costs of conversion, we are interested in clustering our user groups. We cluster the data with k-means clustering, and we select the best cluster number based on the Silhouette score, which measures the degree of separation between clusters. We then train an Optimal Classification Tree (OCT) that maps all the user groups to assigned clusters, which finds an interpretable model that explains the cluster labels using the features. This process will provide transparency and insight into the clustering process and result and will help us to understand the key factors behind different conversion costs.

Results and Discussions

1. Results for Prediction on cost per conversion

We summarize the test results of the *Mean Squared Error (MSE)* and the *Root Mean Square Error (RMSE)* of the four models below. For Campaign No.936, the best RMSE is obtained at 3.226 using the *XGBoost* model. For Campaign No.1178, the best RMSE is obtained at 5.756 using the Holistic Regression model.

	Campaign No.936		**Campaign No.1178**	
	MSE	RMSE	MSE	RMSE
Lasso Regression	132.85	11.526	4339.20	65.872
Holistic Regression	204.74	14.309	33.13	**5.756**
ORT-L	70.74	8.411	288.74	16.992
XGBoost	10.41	**3.226**	343.73	18.539

Table 1. Model Performance of Each Model

We visualize the RMSE of different models against the ground truth y distribution. The RMSE metric can be interpreted in the context of our application, as the degree to which our predictions are different from the true cost of conversions. For Campaign 936, the XGboost model has an RMSE of 3,226, meaning that on average, the predicted cost is about $3 different from the true costs. The ORT-L model predicted ~$8 away from the true costs and the Lasso and Holistic Regression predictions are off by $11.5 and $14 respectively. Similarly, for campaign 1178, the Holistic regression is off by ~$3, and the ORT-L is off by $17. Overall, we can see that ORT-L and XGBoost both provide a fairly accurate prediction.

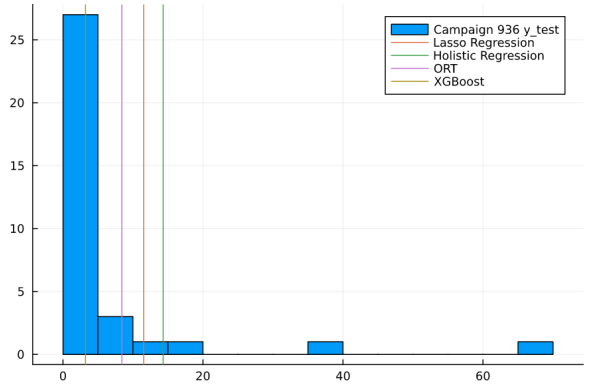


Figure 2. RMSE of Campaign 936 of each model

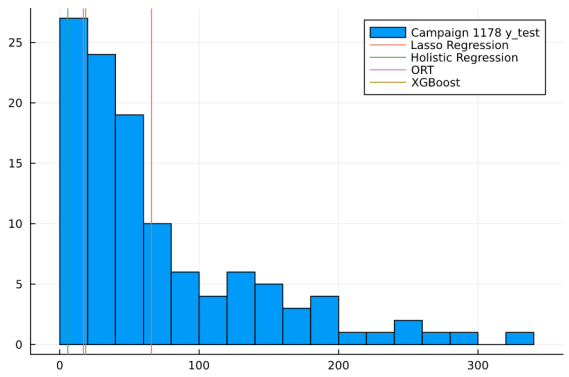


Figure 3. RMSE of Campaign 1178 of each model

Next, we use the Holistic regression and ORT-L result to interpret the prediction for each campaign. We choose holistic regression for Campaign 936 and ORT-L for Campaign 1178.

Holistic Regression. With the best sparse parameter k = 4, we select the following features for Campaign 936: (1) *ageClickRate*^{2} (2) *ageImpressions*, (3) *genderClickRate*, (4) *ageConversionRate*. All the selected features are interaction terms, where one is quadratic transformation and the others are linear transformations. They make intuitive sense because based on our exploratory data analysis, Campaign 936: has a higher conversion for male users between 35-44, therefore, age and gender are important features. We can see that click, impression, and total conversion all contribute to the cost per conversion (approved conversion), which makes intuitive sense. Click rate plays a bigger role and the click rate of different age groups and gender groups are not highly correlated.

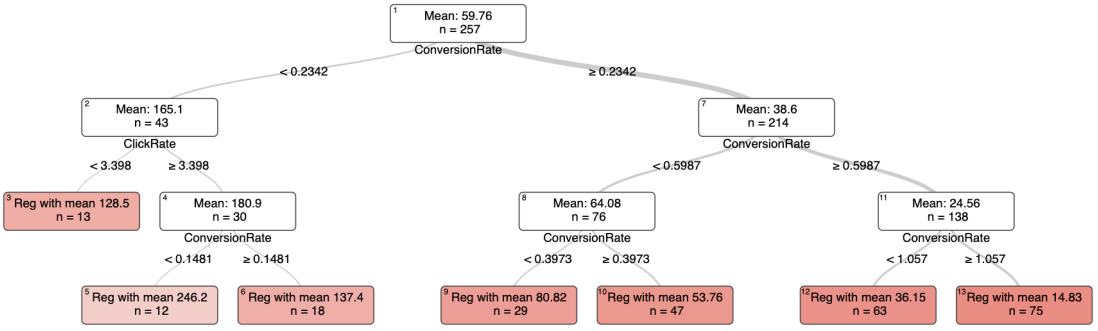


Figure 4. Optimal Regression Tree for predicting Cost of Conversion

ORT-L. The splitting features for the Campaign 1178 model are (1) ConversionRate, and (2) ClickRate, where the Optimal Regression Tree is shown as follows. In general, the Cost of Conversion is low when ConversionRate and ClickRate are high, which aligns with our previous analysis. Specifically, when ConversionRate> 0.0415908% (larger than 60%), the cost per conversion is less than ~$50, when ConversionRate is between 0.0162695% and 0.0415908% (between 23%-60%), the cost is between $50-$80, when ConversionRate is between 0.0102883% and 0.0162695% (between 15%-23%), the cost is between ~$100-200, and when ConversionRate < 0.0102883% (less than 15%), the cost can go as high as ~$250.

Here we also show the Top 5 coefficients of linear models that are fit to the leaves (4 random leaves). We can see that ClickRate is positively correlated with the cost per conversion, and other user characteristics such as age, gender, and interest are also important for prediction.

Variable	Coefficient
ClickRate	2.936
ConversionRate	-2.426
interest_7	0.1307
interest_25	-0.5871
interest_30	-0.4627

Variable	Coefficient
ClickRate	6.371
age_32.0	0.4455
age_47.0	-0.02321
interest_2	-0.3285
interest_10	-0.1462

Variable	Coefficient
ClickRate	6.301
interest_25	-1.82
interest_27	-1.143
interest_107	1.641
interest_110	2.301

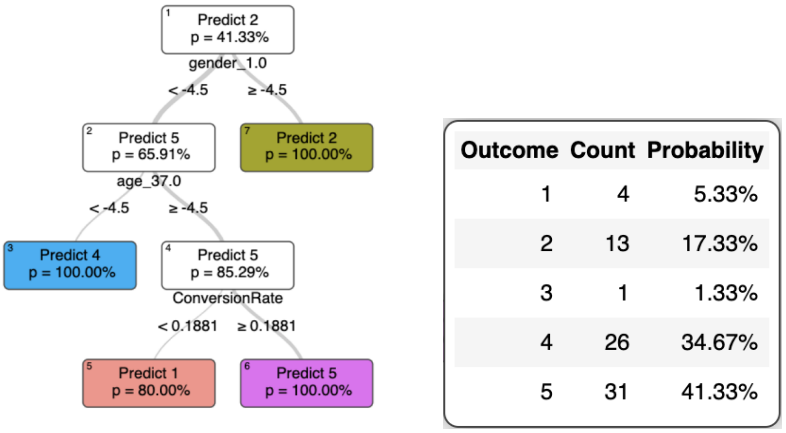
Variable	Coefficient
age_32.0	-0.7598
age_37.0	0.4187
gender_0.0	0.4154
interest_15	1.17
interest_18	1.799

Figure 5. Coefficients of ORT-L Linear Models

2. Customer Segmentation with Interpretable K-Means Clustering.
We analyze Campaign 936 with Interpretable KNN, where we achieve a high Silhouette score when we have 2 and 5 clusters. Given a large user base, we select 5 clusters to allow for more representations of the user dynamics. From the following figure, we conclude that conversion rate, gender = 0, and age = 37 are the branching features.

The user characteristics for each group are as follows.

- Group 1 (cost: $44.28): males not aged between 35-39 with conversion rate < 0.013067%
- Group 2 (cost: $4.4): all females
- Group 3 (cost: $176.4): Unknown
- Group 4 (cost: $5.8): males aged between 35-39
- Group 5 (cost: $2.5): males not aged between 35-39 with conversion rate >= 0.013067%



Outcome	Count	Probability
1	4	5.33%
2	13	17.33%
3	1	1.33%
4	26	34.67%
5	31	41.33%

Figure 3. Interpretable KNN with 5 clusters.

Thus, to obtain a low cost per conversion, our ideal user groups for Campaign 936 are Groups 2, 4, and 5, which are the female user group and males between 35-39, and other groups with high conversion rates. These three groups have a cost of conversion of less than $6, which are significantly lower than the other groups. If we have additional information on the Ads content, it would be interesting to further analyze the relationship between the product and its target user group, as well as the interest listed on users' Facebook profiles.

3. Cost Reduction
In this section, we briefly calculate the cost reduction amount following our selection of the ideal user groups. The original cost to obtain the 118 conversions (purchase actions) is 982.95. Using the user segmentation based on our CTA prediction and KNN clusterings, we propose the following marketing strategy and analyze its cost under two cases.

Best case. We assign all campaigns to Group 5 people because they have the least cost per conversion, which means that we can spend the least money to attain the most purchases. Under this scenario, we achieve a cost of 118 conversions * 2.503 cost per conversion = 120.503, which reduces the original total cost by 89%. One of the problems with this method is that in reality, it might be hard to find a large number of users with the same

characteristics as the Group 5 people. Therefore, we consider this case as the ideal case for cost reduction, which can be feasible when the Ads are targeting a small, specific population.

A realistic case. For a more realistic user assignment, we assign the campaigns based on the real distribution of different groups among all the users. We formulate the cost for each group as 118 conversions * percentage of Group X(1,2,3,4,5) among users * cost per conversion of Group X. Therefore, the total cost is: 118 * 5.33% * 44.2775 + 118 * 17.33% * 4.7314 + 118 * 34.67% * 5.78519 + 118 * 41.33% * 2.50371 = 734.01, which is a reduction of 25.3% for the total cost.

Thus we can conclude that by using our user segmentation groups, we can reduce 25%-89% of advertisement cost while obtaining the same amount of conversions (product purchases).

Conclusions

In this project, we use a Kaggle dataset with information on Facebook advertising campaigns to optimize its marketing strategy. We implemented and compared four models - *Lasso Regression, Holistic Regression, Optimal Regression Trees with Linear Predictions (ORT-L)*, and *XGBoost* for the prediction task. To understand the predictors of conversion cost, we used interpretable clustering with *KMeans Clustering* and *ORT-L*. Using our prescription method, we can decrease the total spent by 89% (ideal scenario) and 25% (average scenario).

Limitations and Future Work

There are two main limitations in our implementation: data and modeling limitations. For data, to protect user privacy, we only have fuzzy data of age and interest, which makes the interpretation of each cluster of users less convincing and informative. Moreover, as aggregated data, it's not possible to conduct prescriptive analysis on an individual user level, which makes our work mainly focus on group-level marketing advertisement placement. However, in reality, it is more common to place ads on an individual level.

For the modeling limitation, we did not include the Optimal Policy Tree for prescription analysis since our data is aggregated on the user groups. For future work, it is possible to consider the prescription framework and apply our findings to help marketing campaigns reduce costs and find their ideal user group.

Contributions

We distributed our work evenly as follows. All members collaborated on project scoping, data preprocessing, results analyzing, and preparing the project report and presentation slides. The prediction models are constructed by the following members:

- Xinyao Han: Lasso Regression, Holistic Regression
- Claire Guan: ORT-L, XGboost, Interpretable Clustering

Appendix

- Dataset: https://www.kaggle.com/datasets/loveall/clicks-conversion-tracking
- Holistic Regression Formulation:

$$
\begin{aligned}
\min_{\beta} \quad & \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \\
\text{s.t.} \quad & -Mz_i \le \beta_i \le Mz_i, \qquad i = 1, \cdots, \tilde{p} \\
& \sum_{i=1}^{\tilde{p}} z_i \le k \\
& \sum_{i:\tilde{X}_i \in \mathcal{T}_j} z_i \le 1, \quad j = 1, \cdots, p \\
& z_i + z_j \le 1, \quad \forall i, j \in \mathcal{HC}(\tilde{\boldsymbol{X}}) \\
& z_i \in \{0, 1\}, \quad i = 1, \cdots, \tilde{p}
\end{aligned}
$$

- ORT-L Formulation:

$$
\begin{aligned}
\min \quad & \frac{1}{\hat{L}} \sum_{i=1}^{n} L_i + \alpha \cdot C + \lambda \sum_{t \in \mathcal{T}_L} \sum_{j=1}^{p} r_{jt} && && (11.1)\\
\text{s.t.} \quad & L_i \ge (f_i - y_i)^2, && \forall i \in [n], \\
& f_i - (\boldsymbol{\beta}_t^\top \mathbf{x}_i + \beta_{0t}) \ge -M_f(1 - z_{ik}), && \forall i \in [n],\ t \in \mathcal{T}_L, \\
& f_i - (\boldsymbol{\beta}_t^\top \mathbf{x}_i + \beta_{0t}) \le +M_f(1 - z_{ik}), && \forall i \in [n],\ t \in \mathcal{T}_L, \\
& -M_r r_{jt} \le \beta_{jt} \le M_r r_{jt}, && \forall j \in [p],\ \forall t \in \mathcal{T}_L \\
& C = \sum_{t \in \mathcal{T}_B} d_t, \\
& \mathbf{a}_m^\top \mathbf{x}_i \ge b_m - (1 - z_{it}), && \forall i \in [n],\ t \in \mathcal{T}_L,\ m \in \mathcal{R}(t), \\
& \mathbf{a}_m^\top (\mathbf{x}_i + \boldsymbol{\epsilon} - \epsilon_{\min}) + \epsilon_{\min} \le \\
& \qquad b_m + (1 + \epsilon_{\max})(1 - z_{it}), && \forall i \in [n],\ t \in \mathcal{T}_L,\ m \in \mathcal{L}(t), \\
& \sum_{t \in \mathcal{T}_L} z_{it} = 1, && \forall i \in [n], \\
& z_{it} \le l_t, && \forall t \in \mathcal{T}_L, \\
& \sum_{i=1}^{n} z_{it} \ge N_{\min} l_t, && \forall t \in \mathcal{T}_L, \\
& \sum_{j=1}^{p} a_{jt} = d_t, && \forall t \in \mathcal{T}_B, \\
& 0 \le b_t \le d_t, && \forall t \in \mathcal{T}_B, \\
& d_t \le d_{p(t)}, && \forall t \in \mathcal{T}_B \setminus \{1\}, \\
& z_{it}, l_t \in \{0, 1\}, && \forall i \in [n],\ k \in [K],\ t \in \mathcal{T}_L, \\
& r_{jt} \in \{0, 1\}, && \forall j \in [p],\ t \in \mathcal{T}_L, \\
& a_{jt}, d_t \in \{0, 1\}, && \forall j \in [p],\ t \in \mathcal{T}_B.
\end{aligned}
$$