Machine Learning 671 - Final Project

Team 5

Aymar Donald Shemeza

Chris Lin

Dylan Liu

Laurie Ye

A. Executive Summary

This report presents a logistic regression model for predicting advertisement clicks, analyzed from a dataset of 13 million entries. The key procedures of data analysis encompasses the following details:

1) Data Transformation: We transformed the ‘hour’ column into two separate columns: ‘day_of_week’ and ‘hour_of_day’. This modification effectively eliminates redundancy in year and month information, while also reducing potential noise in our model's predictions. By isolating the day of the week and the specific hour of the day, we enhance the dataset's relevance and precision, enabling a more accurate and noise-free analysis. We observed that including only ‘day_of_week’ and ‘hour_of_day’ in our model significantly impacts the results, underscoring the importance of these time-related variables in driving meaningful differences in the predictions.
2) Categorical Variable Encoding: we employed Pareto charts to analyze the categorical variables in our dataset, classifying them into two distinct categories based on their frequency distributions: low-cardinality variables (such as C1, C15, banner_pos, site_category, app_category, device_type, and device_conn_type) and high-cardinality variables (including C14, C16, C17, C18, C19, C20, C21, site_domain, app_domain, and device_model). This classification is pivotal in understanding the distinct nature of high and low cardinality variables. Low-cardinality variables, characterized by fewer unique values, are more effectively encoded using one-hot encoding. This technique transforms each category into a new binary column, ensuring simplicity and interpretability for machine learning models. On the other hand, high-cardinality variables, with their vast array of unique values, are better suited for frequency

encoding. This approach replaces categorical values with their corresponding frequencies, efficiently condensing information and reducing the computational burden. By tailoring our encoding techniques to the cardinality of the variables, we effectively transform them into a format that is both machine-readable and optimized for advanced analytical processing.

3) Variable Selection: we streamlined our dataset by eliminating variables exhibiting perfect correlation, notably all ID-related columns, guided by insights from a correlation heatmap analysis. We discarded the ‘id’ column, recognizing that as a mere identifier, it holds no statistical significance in our predictive modeling. To further reduce the dataset's dimensionality, enhancing computational efficiency, we employed a backward elimination approach. This method systematically removed the least significant variables based on their impact on the model. The final step in our variable selection process involved the careful application of p-values, ensuring that only statistically significant variables were retained for the most accurate and efficient predictive analysis.

With all of the steps being proceeded, we are able to construct a logistic regression model which generates a log loss of 0.426 on the validation dataset.

B. Introduction

In the rapidly evolving landscape of digital marketing, the precise prediction of advertisement click-through rates (CTR) has emerged as a critical business imperative. Such predictions empower enterprises to strategically optimize resources and refine user engagement tactics. Our project is at the forefront of this trend, engaging in a deep analysis of intricate patterns hidden within the extensive dataset ‘ProjectTrainingData.csv’, meticulously gathered from

October 21 to October 29. This rich dataset encompasses 24 categorical variables, with the dependent variable 'click' serving as a pivotal indicator of user interaction with advertisements. The remaining 23 independent variables offer a granular view of user-ad dynamics, encompassing a broad spectrum of factors such as user device specifications, ad formats, and contextual elements. These variables provide invaluable insights that not only transcend basic data interpretation but also furnish advanced foresight into the nuances of model construction.

In this archetypal binary classification machine learning endeavor, our primary goal is to engineer a predictive model that meticulously deciphers the intricate interplay between various ad characteristics and user response, specifically the propensity to click. This model aims to achieve a high degree of accuracy, with a particular focus on minimizing log loss, a critical metric in model evaluation. By efficiently reducing log loss, the model not only ensures high predictive accuracy but also aligns seamlessly with the pragmatic demands of real-world applications. Moreover, this model will integrate advanced machine learning techniques to handle the complexities of the dataset, applying state-of-the-art algorithms for feature selection, model tuning, and validation. The objective is to create a robust and scalable solution that can adapt to evolving market trends and user behaviors, thereby enabling businesses to make data-driven decisions that significantly enhance the effectiveness of their digital advertising campaigns.

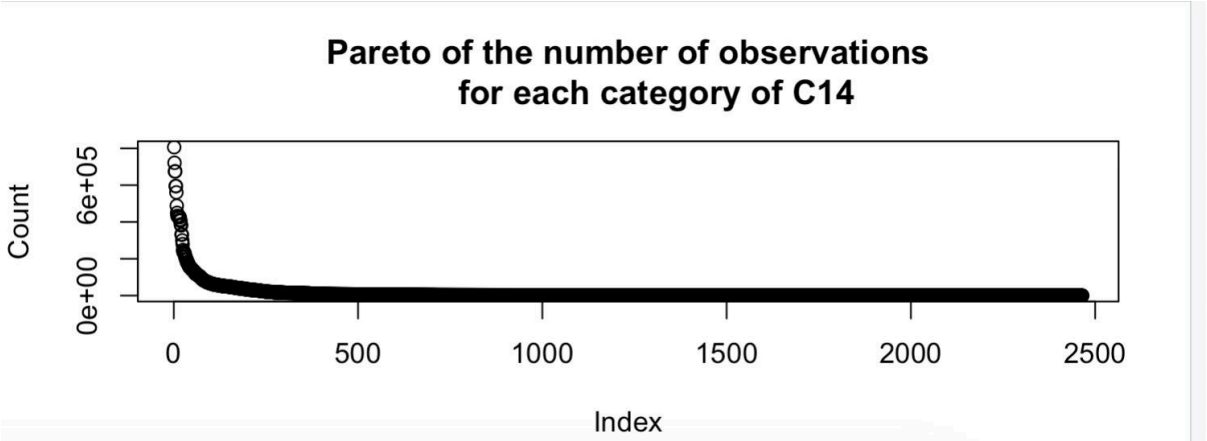
C. Data Processing Process

I. Data Exploration: Unique Count & Pareto Chart

Figure 1. Unique Count Output of Variables

id	click	hour	C1	banner_pos
31991090	2	216	7	7
site_id	site_domain	site_category	app_id	app_domain
4581	7341	26	8088	526
app_category	device_id	device_ip	device_model	device_type
36	2296165	5762925	8058	5
device_conn_type	C14	C15	C16	C17
4	2465	8	9	407
C18	C19	C20	C21	
4	66	171	55	

Figure 2. One of the Pareto Chart Diagrams (C14)



Combining the result from count function (figure 1) with the skewness in the pareto chart distribution (figure 2), we are able to classify our independent variables into two categories. In the later stage, the classification of variable categories will help with encoding:

Low-cardinality-variable: having less number of unique count while more condensed distribution in the pareto chart, meaning that the variable has fewer categories

- C1 & C15: anonymized categorical variables
- banner_pro: the position in the banner
- site_category: a code for the site's category
- app_category: a code for the ad's category
- device_type: the type of the device

- ➢ device_conn_type: the type of the device's connection

High-cardinality-variable: having large number of unique count while strongly right-skewed distribution in the pareto chart, suggesting many variable categories

- ➢ id: the unique identifier of the ad
- ➢ site_domain: an identifier for the site domain
- ➢ app_domain: an identifier for the ad domain
- ➢ device_model: = the model of the device
- ➢ C14 - 21 (excluding 15): anonymized categorical variables

II. Feature Engineering

Transformation of the "hour" Column

In the realm of online advertisement analytics, temporal dynamics are pivotal in understanding user engagement patterns. We refined the temporal granularity of our dataset by extracting 'day_of_week' and 'hour_of_day' from the 'hour' column, which was in the YYMMDDHH format:

- day_of_week: represent specific weekday when ad is posted
- hour_of_day: represent specific hour when ad is posted

This strategic decomposition was driven by the realization that user interaction with advertisements is not only date-specific but also varies significantly throughout the week and the day. By converting these elements into two different variables, there is a higher likelihood to capture the cyclicality of user behavior. This transformation eradicates the redundancy of the year and month information (since all ads were posted in the same year & same month),

which is consistent across our dataset and prevents model obfuscation, thereby enhancing the purity of the time variables.

Encoding Techniques

Referring back to the data exploration output, it is essential to convert the information of each categorical variable into machine-readable format according to their distributions:

One-hot Encoding for Low Cardinality Variables: Our encoding methodology for categorical variables with a limited set of unique values involved one-hot encoding. This technique is instrumental in translating these categories into a binary matrix, enhancing the interpretability of these variables for machine learning algorithms without inflating the feature space.

Frequency Encoding for High Cardinality Variables: For variables characterized by a vast array of categories, frequency encoding was employed. This approach succinctly summarizes the information content of these variables by encoding categories based on their occurrence frequency, thereby maintaining the variable's predictive utility while controlling for computational tractability.

D. Model Building

I. Model Selection: Logistic Regression

Logistic regression model is used in this project for several reasons:

1) Binary Outcome: Logistic regression is specifically designed for binary classification problems. The dependent variable in this dataset, 'click', is binary, where 1 indicates a click and 0 indicates no click. Logistic regression models the probability that each advertisement receives a click, which aligns perfectly with the project's objective.

2) Scalability: Logistic regression's efficiency is crucial for handling our dataset of 30 million rows. Its capability of dealing with large datasets gives us enough room to pivot the model in detail and hence improve its performance quickly.
3) Categorical Variables Handling: Logistic regression can easily incorporate categorical variables through encoding techniques such as one-hot encoding for low cardinality variables and frequency or mean encoding for high cardinality variables. This is essential since all independent variables in this dataset are categorical.

II. Feature Selection: Backward Selection & P-Value

The backward elimination approach is implemented into logistic regression modeling (see figure 3), initiating the analysis with a comprehensive model that includes the full set of variables and gradually dropping the less informative ones: this approach method scrutinizes the contribution of each predictor, sequentially removing the least significant variables to streamline the model. The strength of backward elimination lies in its thorough evaluation of each variable's potential impact when considered in tandem with others, safeguarding against the premature exclusion of variables that may be significant only in the presence of certain combinations. This iterative refinement ensures that the final model is both efficient and retains only those predictors that offer a meaningful contribution to the outcome

Figure 3. Summary Statistics of Backward Elimination (one of the steps)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7022  -0.6465  -0.5410  -0.3306   3.2284

Coefficients: (6 not defined because of singularities)
                       Estimate Std. Error    z value Pr(>|z|)
(Intercept)          -4.849e+09  7.429e+09 -6.530e-01 0.513968
banner_pos1           2.777e-01  1.633e-02  1.701e+01  < 2e-16
banner_pos2           7.653e-01  3.397e-01  2.253e+00 0.024245
banner_pos3           8.541e-01  6.644e-01  1.285e+00 0.198660
banner_pos4           7.615e-01  2.634e-01  2.891e+00 0.003836
banner_pos5           1.745e+00  9.683e-01  1.803e+00 0.071439
banner_pos7           1.740e+00  1.688e-01  1.031e+01  < 2e-16
C11002                2.117e+00  8.412e-01  2.517e+00 0.011844
C11005                1.196e+00  8.012e-01  1.492e+00 0.135620
C11007                5.002e-01  8.808e-01  5.680e-01 0.570125
C11008                       NA         NA         NA       NA
C11010                5.507e-01  8.010e-01  6.880e-01 0.491737
C11012                8.160e-01  8.158e-01  1.000e+00 0.317155
site_category28905ebd  1.700e+00  3.886e-01  4.375e+00 1.21e-05
site_category335d28a8  3.338e-01  4.042e-01  8.260e-01 0.408950
site_category3e814130  1.354e+00  3.887e-01  3.482e+00 0.000497
site_category42a36e14  2.253e+00  5.814e-01  3.875e+00 0.000107
site_category50e219e0  7.136e-01  4.647e-01  1.536e+00 0.124623
site_category5378d028 -1.458e+01  1.603e+03 -9.000e-03 0.992745
site_category70fb0e29  1.171e+00  4.810e-01  2.434e+00 0.014940
site_category72722551  3.996e-01  4.712e-01  8.480e-01 0.396499
site_category75fa27f6  1.037e+00  3.982e-01  2.604e+00 0.009205
site_category76b2941d -7.915e-01  4.395e-01 -1.801e+00 0.071726
site_category8fd0aea4  1.282e-02  1.092e+00  1.200e-02 0.990636
site_category9ccfa2ea -9.030e+03  3.355e+07  0.000e+00 0.999785
site_categorya818d37a -4.504e+15  1.399e+07 -3.218e+08  < 2e-16
site_categorybcf865d9 -4.504e+15  2.122e+07 -2.122e+08  < 2e-16
site_categoryc0dd3be3  9.378e-01  4.299e-01  2.182e+00 0.029132
site_categorydedf689d  2.691e+00  4.345e-01  6.192e+00 5.94e-10
site_categorye787de0e  3.807e-01  1.116e+00  3.410e-01 0.733020
site_categoryf028772b  1.297e+00  3.882e-01  3.341e+00 0.000835
site_categoryf66779e6 -1.197e-01  4.058e-01 -2.950e-01 0.767916
app_category09481d60   8.804e-01  2.827e-01  3.115e+00 0.001842
app_category0bfbc358  -2.464e+01  2.014e+05  0.000e+00 0.999902
app_category0f2161f8   4.926e-01  2.557e-01  1.927e+00 0.054011
app_category0f9a328c   1.547e+00  4.146e-01  3.731e+00 0.000191
app_category2281a340   1.010e+00  1.333e+00  7.580e-01 0.448584
app_category2fc4f2aa  -1.104e+01  3.939e+02 -2.800e-02 0.977637
app_category4681bb9d   1.017e+00  5.244e-01  1.939e+00 0.052552
app_category4ce2e9fc   9.972e-01  3.327e-01  2.998e+00 0.002720
app_category5326cf99  -1.401e+01  5.904e+03 -2.000e-03 0.998106
app_category7113d72a  -1.921e+01  4.067e+03 -5.000e-03 0.996231
app_category75d80bbe   1.267e-01  3.310e-01  3.830e-01 0.701897
app_category79f0b860   2.556e+00  1.428e+00  1.790e+00 0.073462
app_category879c24eb   6.485e-01  3.843e-01  1.688e+00 0.091507
app_category8ded1f7a   3.018e-01  2.576e-01  1.171e+00 0.241448
app_category8df2e842   1.756e+00  7.429e-01  2.364e+00 0.018078
app_categorya3c42688   1.559e+00  5.864e-01  2.659e+00 0.007840
app_categorya7fd01ec   2.446e+00  1.304e+00  1.877e+00 0.060555
app_categorya86a3e89   9.463e-01  6.757e-01  1.401e+00 0.161348
app_categorybf8ac856  -4.504e+15  6.711e+07 -6.711e+07  < 2e-16
app_categorycef3e649   3.641e-01  2.570e-01  1.417e+00 0.156598
app_categoryd1327cf5   5.466e-01  2.727e-01  2.005e+00 0.044999
app_categorydc97ec06   4.248e-01  2.927e-01  1.451e+00 0.146736
app_categoryf95efa07   1.158e+00  2.571e-01  4.505e+00 6.65e-06
app_categoryfc6fa53d  -8.409e-02  4.152e-01 -2.030e-01 0.839489
device_conn_type2     -4.976e-02  2.003e-02 -2.485e+00 0.012966
device_conn_type3     -5.594e-01  4.459e-02 -1.255e+01  < 2e-16
device_conn_type5     -1.164e+00  2.574e-01 -4.525e+00 6.04e-06
C15120                 4.849e+09  7.429e+09  6.530e-01 0.513968
C15216                 4.849e+09  7.429e+09  6.530e-01 0.513968
C15300                 4.849e+09  7.429e+09  6.530e-01 0.513968
C15320                 4.849e+09  7.429e+09  6.530e-01 0.513968
C15480                 4.849e+09  7.429e+09  6.530e-01 0.513968
C15728                 4.849e+09  7.429e+09  6.530e-01 0.513968
C15768                 4.849e+09  7.429e+09  6.530e-01 0.513968
C1620                         NA         NA         NA       NA
C16250                 1.246e+00  5.350e-02  2.330e+01  < 2e-16
C16320                        NA         NA         NA       NA
C1636                         NA         NA         NA       NA
C16480                 4.471e-01  1.130e-01  3.957e+00 7.60e-05
C1650                         NA         NA         NA       NA
C16768                 4.849e+09  7.429e+09  6.530e-01 0.513968
C1690                         NA         NA         NA       NA
C181                  -1.255e+00  4.289e-02 -2.927e+01  < 2e-16
C182                   5.427e-01  1.370e-02  3.962e+01  < 2e-16
C183                   1.461e-02  1.253e-02  1.165e+00 0.243867
```

By employing backward elimination, our model now includes only those variables that meaningfully contribute to the prediction of outcomes, thereby enhancing the model's overall predictive accuracy and interpretability. This selective approach ensures an optimal balance between model complexity and performance, enabling us to craft a robust predictive tool that is both efficient and insightful. This refinement process not only streamlines the feature set but also amplifies the model's predictive prowess in a focused and resourceful manner.

After initially employing backward elimination, we proceeded to utilize P-value selection for the final construction of the logistic regression model. Utilizing a combination of methods proved to be beneficial. The process began with backward selection, which effectively narrowed down a broad set of variables. Subsequently, we fine-tuned our selection based on

P-values, ensuring a more precise and effective model. These are the significant variables we used for logistic regression:

figure 4. P-Value Selection for Variable Significance in Model Analysis

```
> ncol(train_data)
[1] 27
> colnames(train_data)
 [1] "banner_pos0"      "banner_pos1"      "banner_pos2"      "banner_pos3"      "banner_pos4"      "banner_pos5"      "banner_pos7"      "C11002"
 [9] "C11005"           "C11007"           "C11008"           "C11010"           "C11012"           "device_conn_type2" "device_conn_type3" "device_conn_type5"
[17] "C1620"            "C16250"           "C16320"           "C1636"            "C16480"           "C1650"            "C16768"           "C1690"
[25] "C181"             "C182"             "C183"
```

III. Final Model Construction

Strategic Data Partitioning

Before model construction, the training dataset is split into 90% training set and 10% validation set. The allocation of 90% of the dataset to training is a deliberate strategy to leverage the vast swathes of data for model learning, while the 10% reserved for validation serves as a checkpoint to gauge the model's predictive prowess on an independent data subset. This split strikes a calculated balance, ensuring a comprehensive learning process while maintaining a buffer against overfitting.

Hyperparameter Tuning & Cross Validation

With the selected variables, our logistic regression model is established based on the cross-validation approach, there are two imperatives:

- Cross-validation amplifies the statistical validity of our model evaluation and mitigates the variability inherent in a singular partitioning of the dataset.
- By cycling through different training and validation subsets, cross-validation ensures a comprehensive assessment of the model’s predictive stability.

Furthermore, we tested the model performance on different value of "lambda" parameter to optimize the performance: regularization emerges as a cornerstone in refining logistic regression models, with the parameter lambda serving as the fulcrum for balancing model complexity and predictive accuracy. By penalizing the magnitude of the coefficients, lambda effectively reins in overfitting, ensuring that our model remains agile and generalizable.

Model Evaluation: log loss

We implemented log loss to evaluate the model performance given that this is a highly suitable metric for evaluating models in binary classification tasks like advertisement click rates, log loss provides probabilistic interpretation of the model's predictions:

- log loss penalizes overconfident incorrect predictions more severely, encouraging models not only be accurate but also reflect the true likelihood of an event occurring
- log loss rewards models that offer calibrated probability estimates, its sensitivity to the uncertainty of predictions contributes to measure robustness

In the context of an advertising scenario, where understanding the confidence of a click prediction is as important as the prediction itself, log loss aligns well with business objectives, offering clear insights into model performance

Prediction: Testing Data

After loading the test data, we identified 27 significant dummies variables from banner_pro, C1, C16, C18, and device_conn_type. In this case, we chose to transform these categorical variables into dummy variables for prediction.

E. Summary

In conclusion, this project has developed a logistic regression model that excels in predicting advertisement click-through rates, achieving a log loss of 0.426. This level of precision is the result of thorough data preprocessing, which included advanced feature engineering and strategic variable encoding. We implemented backward elimination to ensure the model retained only the most significant predictors, thus striking a balance between simplicity and predictive power. With its satisfactory log loss, this model is poised to become an invaluable tool for businesses seeking data-driven decision-making capabilities. It enhances targeting precision and improves the efficacy of advertising efforts. By predicting user engagement, organizations can more effectively allocate their marketing resources, engaging with the ideal audience at the opportune moments. This boosts revenue while also improving the user experience. The model transcends its role as a mere predictive mechanism; it acts as a catalyst for smarter, more efficient marketing strategies. Looking ahead, there is potential to further refine our predictive capabilities by exploring other advanced methodologies, such as random forest or gradient boosting techniques.