Advertisement Data Prediction

A Comprehensive Report

TECH-GB-2336 - Data Science for Business : Technical
Section 30
Professor: Dr. Chris Volinsky

Submitted by:
Leon Zhang
Ishan Miglani
Keya Keya
Qing Shen

Group 8
Date : 05/08/2024

NYU | STERN

Content

1. Business Understanding

Meetsocial is a leading Chinese enterprise who specializes in providing overseas marketing services for Chinese businesses. The primary service it provides is to put on online advertisements for mobile apps businesses in collaboration with it in oversea regions. However, the decision-making process for online advertisement is complex and often relies heavily on the personal experience of optimizers. Limited use of historical data is often observed in such processes. The company's clients are categorized into key accounts (KA) and small-to-medium enterprises (SMEs). While KA clients are serviced by dedicated optimization teams, SMEs, which constitute over 80% of the client base, cannot afford such luxuries. These SMEs are confident in their products but lack familiarity with overseas markets and marketing strategies. They are uncertain about the target demographics and appropriate advertising strategies in foreign countries. Therefore, they seek to leverage the company's extensive historical data for strategic guidance. Given their sensitivity to cost and effectiveness, these clients demand immediate results and may discontinue their campaigns if initial outcomes are suboptimal. Thus, something that does not require too much effort and could produce reliable results is needed.

As the data science group in Meetsocial, a predictive model is the perfect tool that comes to our mind immediately after understanding the objectives. To be specific, a reliable predictive model that can forecast the advertising outcomes based on specific budgets and timelines might be crucial for this scenario. Now the business question of interest is clear: how can we develop a predictive model to accurately forecast advertising outcomes for SMEs (particularly focused on

app promotion), ensuring they optimize their campaigns and achieve their goals? More specifically, how can the model aid in determining the best advertising settings and budget allocation to maximize impressions, clicks, and installs?

2. Data Understanding

Before diving into details of the model, it's important to understand the historical data that is to be put into it. The data we have is the comprehensive records of historical advertisements that Meetsocial performed for KA clients. There are two major tables containing 320000 rows and 30000 rows separately. The first table records the methods about how the ad was put on, date, and results of one specific campaign. The second table records more info like targeting demographics about a specific ad that can contain many campaigns. These two tables' records can be connected by the ad's id so that more information about a specific campaign can be learnt. Even though we have massive records, there are many NaN values existing and we believe that more information about the products that are to be campaigned about is needed to better predict the outcome. Other problems also reside in current data, for example, the results of the ad (impressions, clicks, installs, and purchases) are extremely biased. To be specific, the overall spread of the result is not balanced at all, it ranges from 0 to 1000000 and more than 95% of the data are very close to 0. Therefore, when we start to do data preparation, this is something that we have to address for the model to behave normally. The justification of the performance of one ad will also need further discussion since the budget spent on different campaigns differs a lot. As a result, impressions of ads should be strongly correlated with the money spent on it. How to justify an ad's performance requires more attention later.

3. Data Preparation

Targets

Clicks, Installs, Purchases, Impressions, Click to Impression Ratio (CTR)

CTR is not found in our dataset. It is a target that we have generated from the Clicks and Impressions column. This is done because CTR is a very relevant indicator in advertisement performance.

Merging Data

We had 2 excel sheets to combine to get our data. The first sheet with 30,000 rows contains data about ads that stay consistent across multiple days of advertisements. Settings such as campaign_objective, targeting_gender and targeting_age are contained in this table.

The second sheet with 320,000 rows contains data about advertisement performance on a daily basis. That is, how much was spent on an advertisement on a particular day and the targets it generated(clicks, installs,purchases,impressions,etc). There are a few other features about the ads as well such as which device it targeted, the os_type, product_name, ad_network_type, a few identifying numerical features, etc.

For data collected from Google and put into the excel sheets we noticed that across the 2 sheets the primary key to join them was the ad_id column in combination with the dates. The ad calender_date column from sheet 2 should be in between the ad_first_dt and ad_late_dt from sheet 1.

This was not the case for the data collected from Facebook. For Facebook data the ad_id notations were not consistent across both sheets. One was in scientific notation and the other was in normal numeric notation. The unique column here was product_id which was in string format. The combination for the dates was the same as was done for the Google data.

Data Cleaning

After combining the data as described above we were left with data with features from both tables and each row corresponding to performance of an ad for a single day. Some features which were completely empty such as gender and product_category were deleted.

For a few features which had a few null values, we did not want to lose out on the other features of these rows. So to handle that we added values manually. This was done for 'targeting_gender' which if null we set to 'All genders'.This was also done for 'campaign_objective' which if null was set to 'Unknown'.

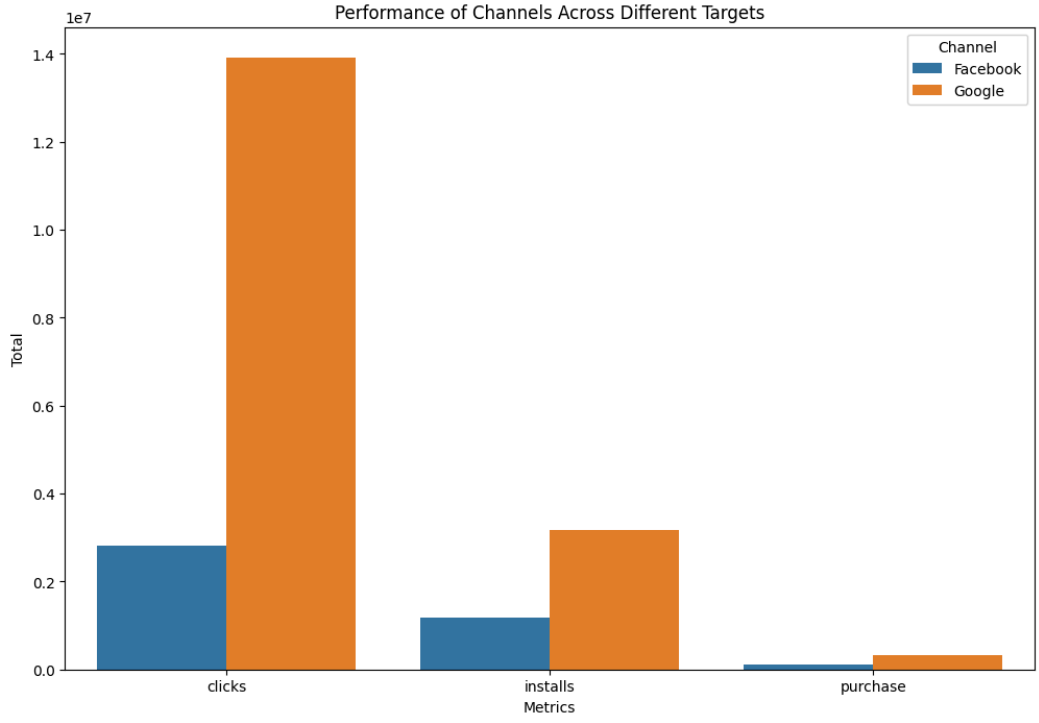
For days where no money was spent on an advertisement no targets were generated at all. This is consistent with advertisement behavior where the ad is not displayed anywhere if no money is spent. These rows were also then deleted from the data as they are not relevant.After this we deleted all rows with null values as we could not collect data for those values and also deleted duplicate rows.

Exploratory Data Analysis

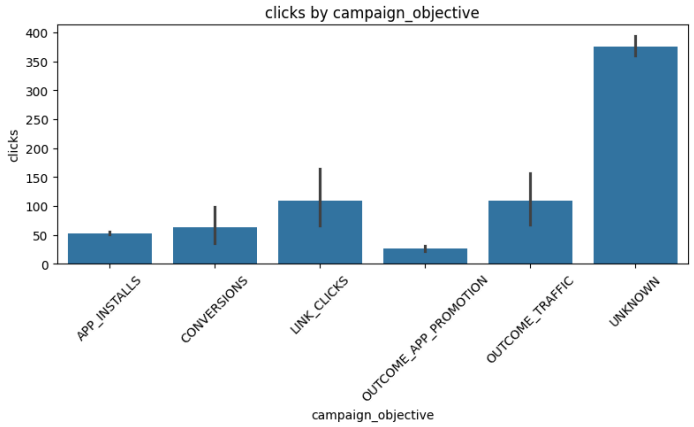
We performed extensive data analysis with making correlation matrices as well as bar graphs between different features and all the targets. A sample code is shown below.

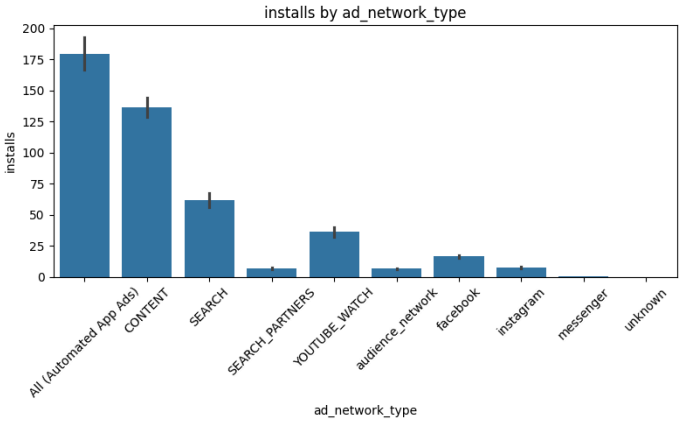
```python
import seaborn as sns
features = ['campaign_objective', 'channel', 'os_type', 'ad_network_type', 'device']
target_variable = 'clicks'

# Function to generate bar graphs
def generate_bar_graphs(df, features, target_variable):
    for feature in features:
        plt.figure(figsize=(8, 5))
        sns.barplot(x=feature, y=target_variable, data=df)
        plt.title(f"{target_variable} by {feature}")
        plt.xlabel(feature)
        plt.ylabel(target_variable)
        plt.xticks(rotation=45)
        plt.tight_layout()
        plt.show()

# Call the function
generate_bar_graphs(df, features, target_variable)
```

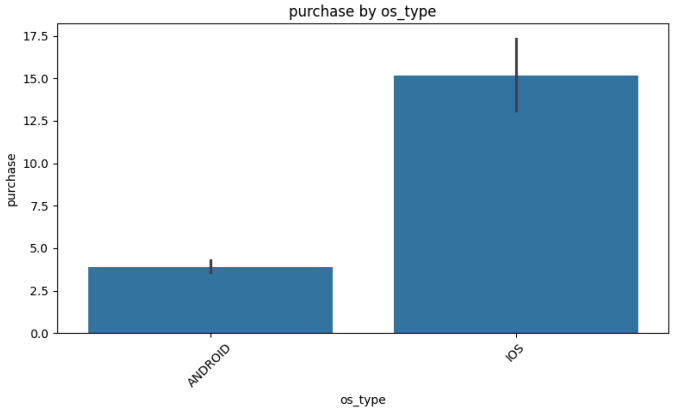
To get a sense of the data a few of the visualizations are shown here. The image below shows the distribution of 3 of our targets - clicks, installs and purchase. These are distributed across the 2 channels we have - Facebook and Google. We see a pattern that is consistent from advertisement campaigns with clicks being much higher than installs and installs in turn being higher than purchase.



The 3 graphs below 3 features - campaign_objective, ad_network_type and os_type against 3 of the targets - clicks, installs and purchase. These show the levels of these 3 categorical features and their distribution across the targets.







Feature Engineering

Since at this point we had 90K rows and 19 columns, 5 of which were identifying information, 5 of them were targets we realized we had a low number of features to train our models on. To address this we decided to gather more features for the data we had. For numerical data we wanted to collect ratings, number of ratings and number of reviews as they might influence how

likely a person is to click on or install an app which we advertise. For categorical data we wanted to collect the various categories the app was put under inside the app store. The category of an app might influence advertisement performance.

We built 2 different web scrapers. One for the google play store and one for the Apple app store.They collected the features mentioned above. To maintain consistent categories from the 2 different app stores we performed some manual sorting as well in the few cases it was required.

After scraping we are left with 15 unique categories with each app having 1 or more of those categories. The image below shows the complete features we scraped.

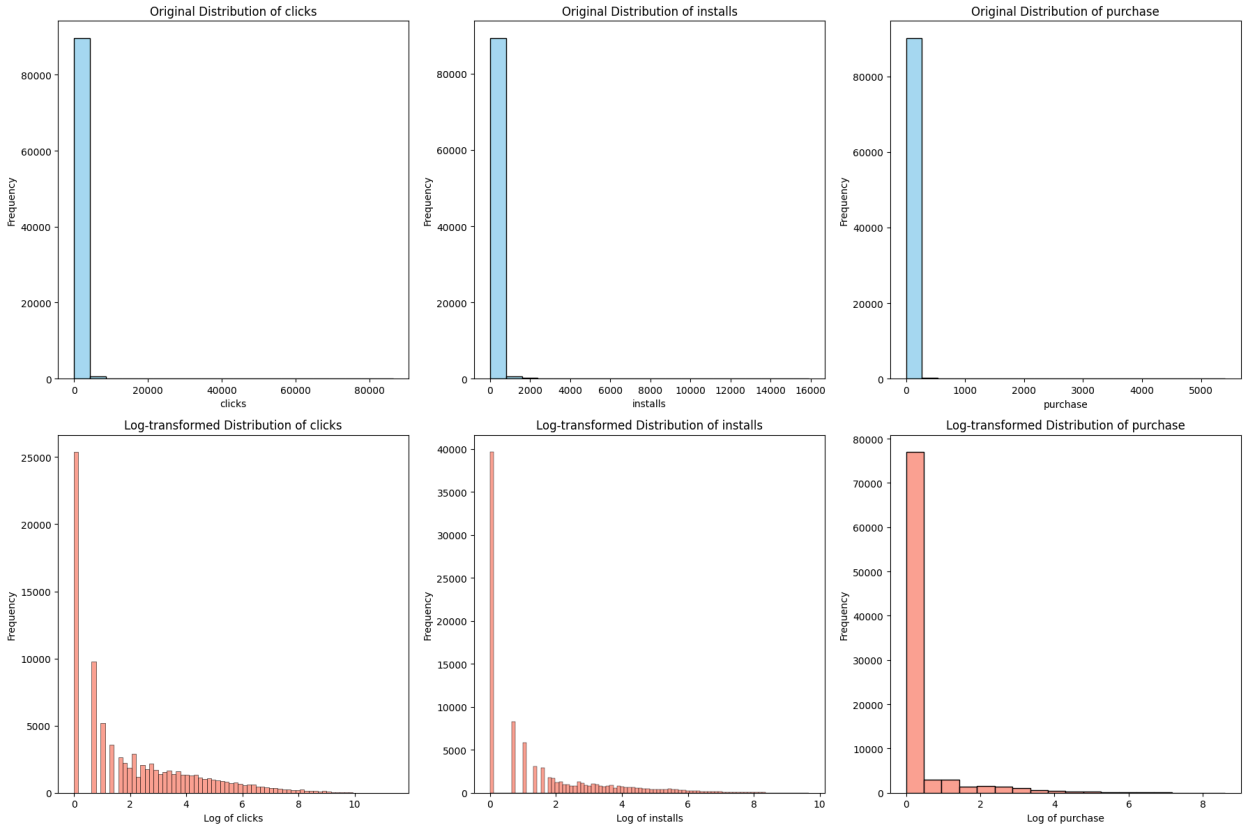
```
20  PUZZLE                  26145 non-null  float64
21  CASUAL                  26145 non-null  float64
22  SINGLE PLAYER           26145 non-null  float64
23  STYLIZED                26145 non-null  float64
24  OFFLINE                 26145 non-null  float64
25  ROLE PLAYING            26145 non-null  float64
26  BOARD                   26145 non-null  float64
27  ABSTRACT STRATEGY       26145 non-null  float64
28  LUDO                    26145 non-null  float64
29  MULTIPLAYER             26145 non-null  float64
30  COMPETITIVE MULTIPLAYER 26145 non-null  float64
31  REALISTIC               26145 non-null  float64
32  CASINO                  26145 non-null  float64
33  SOCIAL                  26145 non-null  float64
34  SIMULATION              26145 non-null  float64
35  Rating                  26145 non-null  float64
36  Num_Ratings             26145 non-null  float64
37  Num_Reviews             26145 non-null  float64
```

We also noticed a discrepancy with the targeting_age feature. It had overlapping ranges for different apps. For example, 13-65,15-65,17-49 were a few of the values found. To address this we created 4 features for each of the 4 ranges: 13-26,26-39,39-52,52-65. We added a 1 or a 0 for each column depending on value in the targeting_age feature. So 13-65 would have 1's in all 4 columns and so on.

Final Data Preparation

After this was done we created dummy variables for all our categorical features and removed all null values again and we were left with a dataset of 26,889 rows and 59 features.

For our target variables we noticed that their distribution was very skewed. They were either many 0's or very high values. To address this, we used the popular method of logarithmic transformations of the target variables to spread them out more evenly and help improve our model's R squared and reduce its root mean squared error. The change in distribution is shown below:



To make sure that no numeric feature dominates over our model due to the different ranges of their values we also used a StandardScaler which standardizes features by removing the mean

and scaling to unit variance, transforming each feature to have a mean of zero and a standard deviation of one. For example rating was from 0 to 5 whereas num_of_reviews could range from 0 to anything. A large num_of_reviews could unfairly influence our model training and predictions.

4. Model Development

Overview

In the model development phase, our primary objective was to identify the best performing machine learning models for predicting various advertising outcomes such as clicks, installs, purchases, and impressions. Given the complexity and variability of the data, multiple regression models were evaluated to identify the most effective approach for our predictive needs.

In the initial phase of our modeling process, we addressed the skewed distribution of our target variables, which often featured heavy tails from outliers, by applying a logarithmic transformation. This technique is crucial for normalizing data distributions, reducing the impact of outliers, and improving the predictive performance of our models. By compressing the data range, logarithmic transformation makes the data more compatible with the assumptions of many statistical learning methods.

Model Training

We split our dataset into an 80% training set and a 20% testing set to enhance the robustness of our model evaluations and prevent overfitting. To normalize the input features, we applied feature scaling using StandardScaler, which adjusts each feature to have zero mean and unit variance. This normalization ensures that no single attribute disproportionately influences the model due to its scale. Both the training and test datasets were transformed using the fitted scaler, ensuring consistency and uniformity in our data throughout the model training and prediction phases.

Framework for Robust Predictive Modeling

Building upon the transformed data, our objective shifted to constructing a robust predictive framework. This framework was designed to be comprehensive and flexible, capable of integrating and adapting to the inherent complexities and variabilities within the data. To achieve this, we explored a variety of machine learning models, each with strengths in handling different aspects of prediction and complexity:

1. **Linear Regression:** Used as a baseline for performance comparison, linear regression provided a straightforward interpretation of the effect of each predictor on the advertising outcomes.
2. **Random Forest and XGBoost:** These methods were particularly advantageous due to their robustness against overfitting and their ability to model non-linear relationships effectively. Both Random Forest and XGBoost incorporate mechanisms to conduct feature selection implicitly, which is crucial in dealing with high-dimensional data.

3. **Support Vector Regression (SVR):** Chosen for its capacity to handle large feature spaces and its effectiveness in high-dimensional environments, SVR was tested for its performance in scenarios where the relationships in the data were not necessarily linear.
4. **Decision Tree Regression:** Provided a more intuitive understanding of the data's structure by splitting the variables at several decision nodes. This model was useful for capturing non-linear interactions between the variables but required careful tuning to avoid overfitting.
5. **Ensemble Model (Combination of RF and XGBoost):** By combining the strengths of Random Forest and XGBoost, this hybrid approach aimed to leverage the individual advantages of each model while mitigating their respective weaknesses.

5. Model Evaluation

Evaluation Metrics

RMSE (Root Mean Square Error) and R^2 (Coefficient of Determination) are chosen as performance metrics because they provide complementary information about the accuracy and goodness of fit of a predictive model, which are crucial for evaluating its effectiveness given the prediction objectives. Here's why they are relevant:

RMSE (Root Mean Square Error):

- **Measures Accuracy:** RMSE provides a clear measure of the model's accuracy by quantifying the average magnitude of the prediction errors — the differences between the values predicted by the model and the observed values.

- **Sensitive to Outliers:** RMSE is particularly sensitive to outliers. This is beneficial if the prediction objective requires a high level of accuracy and we need to penalize large errors more severely.
- **Scale-Dependent:** It is in the same units as the predicted variable, making it interpretable and practical for assessing the model's predictive capability in a real-world context.

R^2 (Coefficient of Determination):

- **Explains Variance:** R^2 indicates how well the data fit a statistical model — or the proportion of variance in the dependent variable that can be predicted from the independent variable(s).
- **Comparative Measure:** It provides a normalized measure that can be used to compare the explanatory power of models on the same task. An R^2 value close to 1 indicates that the model explains a large portion of the variance in the target variable.
- **Modeling Efficiency:** When the prediction objective includes explaining the influence of input variables on the outcome, R^2 helps in understanding the strength of the relationship between the model and the dependent variable.

Together, RMSE and R^2 provide a comprehensive view of model performance. RMSE measures the model's prediction errors, showing how close the predicted values are to the actual data. R^2 indicates the proportion of variance in the response variable explained by the model, highlighting its effectiveness. For our advertising goals, minimizing RMSE ensures accuracy in predicted metrics like clicks and impressions, while maximizing R^2 confirms the model's capacity to capture and reflect campaign performance variability, crucial for informed decision-making.

Why Not Other Metrics?

Mean Absolute Error (MAE): While MAE is another common metric for regression, it does not penalize large errors as heavily as RMSE, making RMSE more appropriate in scenarios where larger errors are more detrimental.

Mean Absolute Percentage Error (MAPE): MAPE can be useful but is less reliable when dealing with zero or near-zero actual values as it leads to undefined or infinite errors. It also inherently gives higher weight to percentage errors in smaller true values, which can be misleading in some contexts.

Adjusted R^2: Adjusted R^2 could be used instead of R^2, especially when the model has a large number of predictors. Adjusted R^2 accounts for the number of predictors in a model, providing a more adjusted measure of the goodness of fit. However, for simplicity regular R^2 might suffice.

Performance with Log-Transformed Data

As evident from the table below, XGBoost model showed a promising balance with an RMSE of 0.778 and an R^2 of 0.899, suggesting it's quite effective at predicting click outcomes while maintaining a low error rate.Moreover, compared to our baseline (Linear Regression), we could see 56% and 88% improvement in metrics, respectively.

Target: Clicks

Model Tested	RMSE	R^2

Linear Regression (Baseline)	1.770	0.478
Random Forest	0.818	0.889
SVR	1.516	0.617
XGBoost	**0.778**	**0.899**
Decision Tree Regression	1.064	0.812
Ensemble Model (RF + XGBoost)	0.799	0.894

Targets	RMSE	R^2
Clicks	0.778	0.899
Installs	0.775	0.860
Purchase	0.756	0.896
Click/Impression (CTR)	0.869	0.738
Impressions	0.903	0.916

Similar to Clicks, our other targets performed better in XGBoost compared to the other models we tested.

Reevaluation with Inverse Transformation

Upon applying the inverse transformation to bring predictions back to their original scale, we observed a shift in performance.Below are the true values for targets - clicks,installs,impressions and purchases.

Target: Clicks

Model Tested	RMSE	R^2
Linear Regression (Baseline)	15166149.138	-77710823.252
Random Forest	**999.881**	**0.662**
SVR	1607.041	0.127
XGBoost	1338.025	0.395
Decision Tree Regression	1488.378	0.252
Ensemble Model (RF + XGBoost)	1096.979	0.593

Target: Installs

Model Tested	RMSE	R^2
Linear Regression (Baseline)	135126.306	-174451.853
Random Forest	**184.732**	**0.674**
SVR	273.346	0.286
XGBoost	192.516	0.646
Decision Tree Regression	243.284	0.434
Ensemble Model (RF + XGBoost)	187.583	0.664

Target: Purchase

Model Tested	RMSE	R^2
Linear Regression (Baseline)	414.261	-5.887
Random Forest	77.503	0.759
SVR	93.083	0.652

XGBoost	**62.449**	**0.843**
Decision Tree Regression	73.191	0.785
Ensemble Model (RF + XGBoost)	96.804	0.624

Target: Impressions

Model Tested	**RMSE**	**R^2**
Linear Regression (Baseline)	11220724	-4188
Random Forest	102952	0.647
SVR	164494	0.099
XGBoost	**90893**	**0.725**
Decision Tree Regression	113178	0.57
Ensemble Model (RF + XGBoost)	94642	0.702

Target: Click/Impression (CTR - Click Through Rate)

Model Tested	**RMSE**	**R^2**
Linear Regression (Baseline)	0.089	0.083
Random Forest	0.090	0.068
SVR	0.110	-0.402
XGBoost	0.089	0.097
Decision Tree Regression	**0.086**	**0.141**
Ensemble Model (RF + XGBoost)	0.091	0.0474

- The effectiveness of XGBoost, which previously excelled, varied with different target variables indicating a need for model tuning or potential overfitting on transformed targets.
- Random Forest showed consistent performance across different metrics, suggesting its reliability and robustness.
- Notably, the Ensemble Model combining Random Forest and XGBoost didn't always lead to improved predictions, challenging our expectation of ensemble methods outperforming individual models.

Challenges and Considerations

The disparity in RMSE and R^2 post-inverse transformation underscores the sensitivity of models to the scale of data. This raises important considerations about the appropriateness of data preprocessing methods like log transformation.

Our analysis reveals that no single model universally excels across all metrics and targets, highlighting the need for targeted model selection based on specific campaign goals.

6. Insights and Business Impact

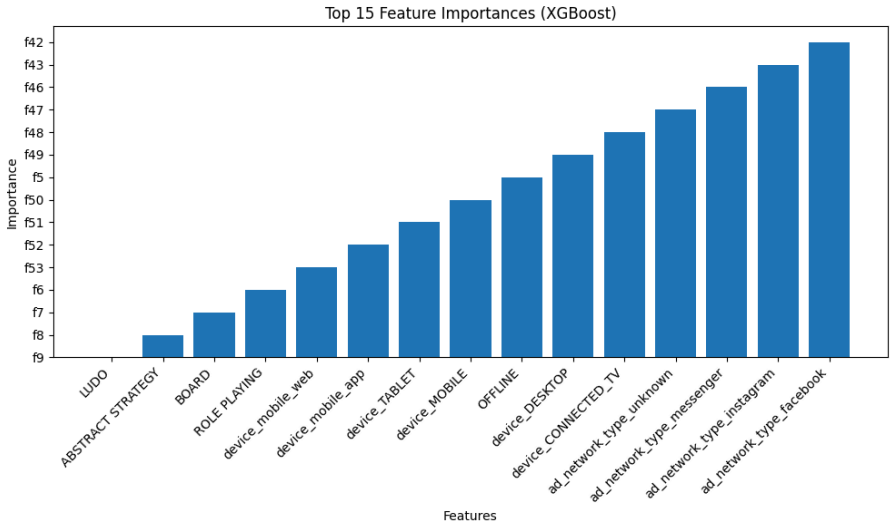
Overview of Feature Importance

Feature importance is a technique used to identify which variables in a dataset are most influential in predicting the target variable. In our model, understanding feature importance has allowed us to discern which factors most significantly affect the performance of online

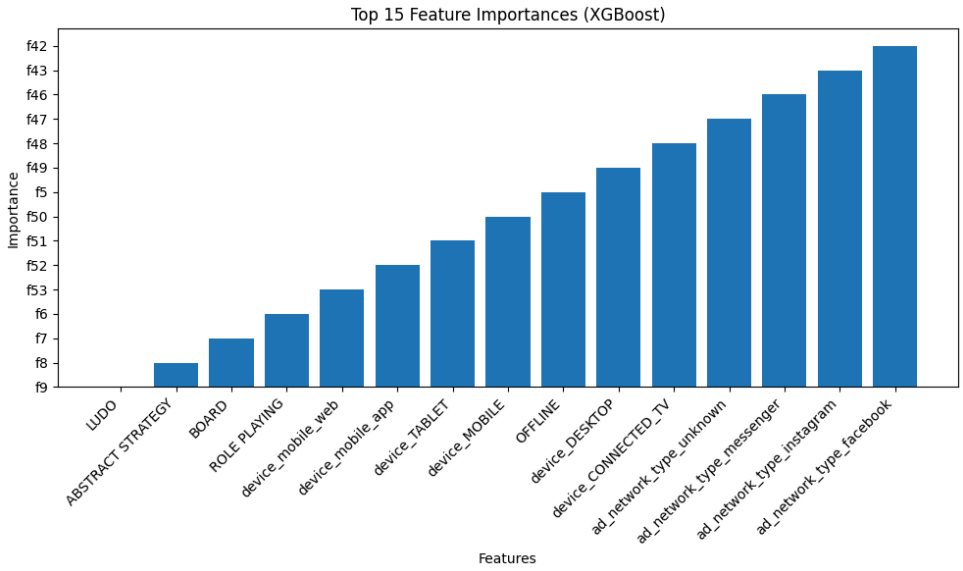
advertising campaigns across different metrics such as clicks, installs, impressions, and purchases.

Below are the graphs for feature importance of our four targets:

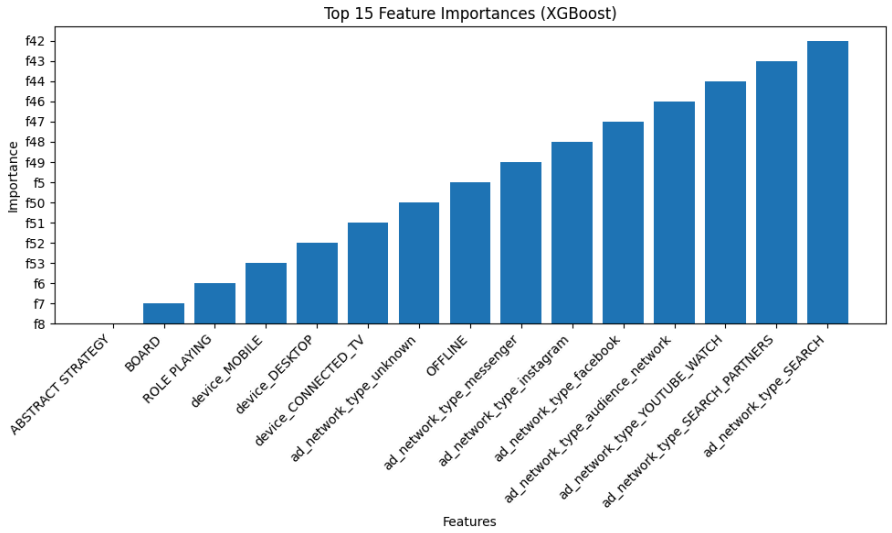
Clicks



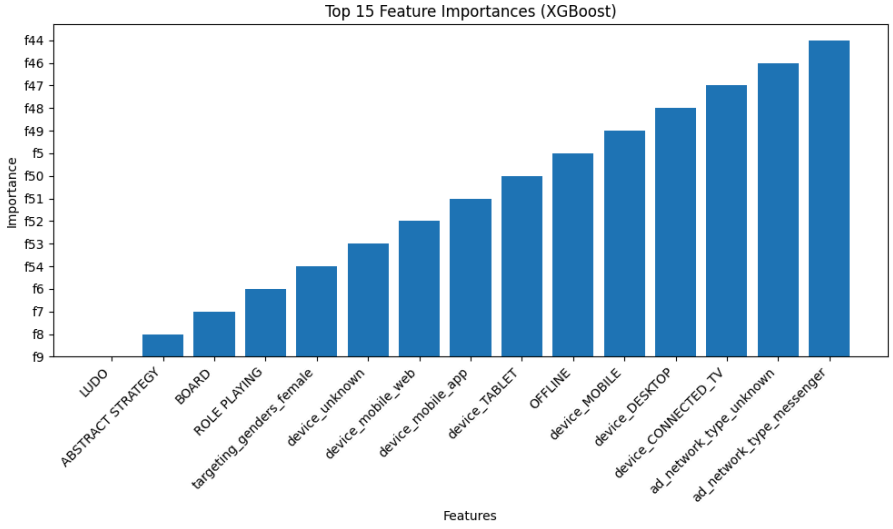
Installs



Purchases



Impressions



The graphs are presented as bar charts, where each bar represents a feature's importance score relative to others in predicting the respective outcome.The length of each bar indicates the weight or contribution of the feature to the model’s predictive accuracy, with longer bars signifying greater importance.These visualizations help stakeholders quickly understand which aspects of their campaigns need emphasis or reevaluation to enhance performance.

Insights:

Mobile and Social Dominance: Our analysis reveals that for objectives like 'Clicks' and 'Installs,' mobile devices and social media platforms, particularly Facebook and Instagram, are dominant drivers of user engagement and conversions. This underscores the critical importance of mobile optimization and targeted social media campaigns in driving meaningful interactions and installations.

Broad Platform Strategy: In contrast, objectives focused on 'Impressions' benefit from a broader approach, with desktops and connected TVs playing significant roles alongside social networks. This highlights the importance of a diversified platform strategy to maximize visibility and ensure ads are effectively displayed across different devices, ultimately reaching a wider audience.

High-Impact Social Platforms: Purchase decisions are strongly influenced by engagement on high-impact social platforms, with platforms like Instagram playing a pivotal role in converting impressions into sales. This underscores the significance of tailoring content specifically for these platforms to facilitate not just interaction but conversion.

Business Impact:

Strategic Focus: Armed with these insights, businesses can strategically focus their advertising efforts on channels and tactics that yield the highest returns. By prioritizing mobile optimization and targeted social media campaigns for click and install objectives, companies can increase user engagement and drive more installations, ultimately boosting brand visibility and market share.

Optimized Resource Allocation: Understanding the importance of a diversified platform strategy for maximizing impressions enables businesses to allocate resources more effectively. By ensuring ads are effectively displayed across various devices and platforms, companies can reach a broader audience and increase brand exposure, ultimately driving greater brand awareness and consideration.

Enhanced Conversion Strategies: Insights into the influence of high-impact social platforms on purchase decisions empower businesses to refine their conversion strategies. By optimizing user experience and tailoring content specifically for these platforms, companies can increase the likelihood of conversion, ultimately driving higher sales and revenue.

In summary, these insights provide businesses with actionable guidance for optimizing their advertising strategies to achieve their objectives more effectively. By leveraging the dominant drivers of user engagement and conversion revealed by our analysis, companies can drive tangible business results and gain a competitive edge in the digital advertising landscape.

7. Deployment Plan

The implementation and deployment of our predictive model represent a critical milestone in our journey towards leveraging data-driven insights to enhance business decision-making. With the model's capabilities validated through rigorous testing and evaluation, the focus now shifts

towards seamlessly integrating it into our existing infrastructure and making its insights readily accessible to stakeholders across the organization.

Integration with Business Systems:

The first phase of deployment involves integrating the predictive model with our existing business systems. To achieve this, we have developed robust APIs that enable seamless communication between the model and other core systems. These APIs have been meticulously designed to ensure compatibility with various data formats and protocols, facilitating smooth data exchange. Furthermore, stringent security measures have been implemented to safeguard data integrity and confidentiality during transmission, aligning with our commitment to data privacy and compliance.

Cloud Deployment:

To handle the scalability requirements of our model and ensure accessibility to a wide range of users, we have opted for cloud deployment. Leveraging a leading cloud platform, we have configured the necessary infrastructure to accommodate the computational demands of the model. By harnessing the elasticity of the cloud, we can dynamically scale resources based on demand fluctuations, ensuring optimal performance even during peak usage periods. Additionally, robust data redundancy and disaster recovery mechanisms have been implemented to mitigate the risk of service interruptions and data loss, bolstering the reliability and resilience of our deployment.

User Interface Development:

Central to our deployment strategy is the development of a user-friendly dashboard that empowers business users to interact with the model's predictions and insights effortlessly. Through extensive user research and iterative design iterations, we have crafted an intuitive dashboard interface that prioritizes usability and clarity. Users can easily customize inputs, visualize data, and derive actionable insights through interactive features such as filters and drill-down capabilities. By incorporating data visualization techniques, such as charts and graphs, we aim to present complex information in a digestible format, enabling stakeholders to make informed decisions with confidence.

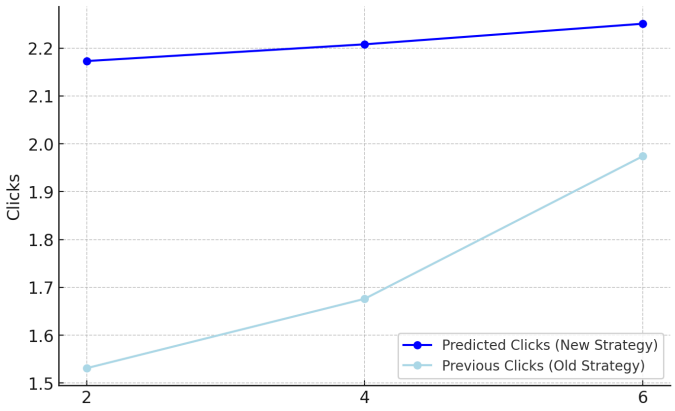
8. Case Study

One of the challenges Meetsocial faced is selecting the optimal advertising strategy for a new client, especially when specific information such as the client's product categories is known. To validate the effectiveness of our model, we conducted a case study on APPClubillion Vegas Casino Slots, a client not included in our training and testing datasets. To evaluate the expected effects of all possible strategy combinations, we generated 23,040 different combinations by multiplying the number of possible choices for each spend level. The formula for the total combinations is as follows:

Total Combinations = (Number of Age Combinations) × (Number of Campaign Objectives) × (Number of Channel Google Options) × (Number of OS Type iOS Options) × (Number of Ad Network Types per Channel Google Option) × (Number of Device Types) × (Number of Gender Combinations)

Additionally, we had access to all advertising data for this client from December 1, 2022, to December 31, 2022. By comparing the historical advertising strategies with the optimal strategy suggested by the model, the specific strategy comparison is detailed in the settings and results for three most commonly used budget levels regarding ad click outcomes.

Settings	New Strategy	Previous Strategy
target_age	13-26	45-65
channel	Facebook	Google, Facebook
os_type	Andriod	Andriod
campaign_objective	CONVERSIONS	APP_INSTALLS
ad_network_type	audience_network	All
device	Connected_TV	mobile_app
targeting_genders	female	female,male



Especially, with a single ad budget of $2 a day, based on the same budget, the updated ad strategy is expected to result in a 41.9% lift in the number of clicks. This case study validates the utility of a data-driven approach in optimizing advertising strategies. By leveraging comprehensive data analysis and strategic testing, digital marketers can significantly enhance the effectiveness of their campaigns.

9. Future Work

Data Collection

In our future data collection part, we will first delve deeper into data integration by synthesizing data across various advertising platforms to obtain a comprehensive view of ad performance and

user interactions. This will also involve implementing systems that can process and analyze data in real time, allowing for dynamic adjustments to advertising strategies based on current trends and user feedback.

Secondly, we plan to advance our analytical capabilities by developing predictive analytics models. These models will forecast user behaviors and preferences, enhancing the precision of ad targeting. Alongside, we will deploy sophisticated machine learning algorithms to automate the identification of optimal ad placements and creative strategies, ensuring maximum engagement and efficiency.

Thirdly, we aim to optimize user experience by refining personalization algorithms to offer more tailored ad experiences, which are crucial for increasing user relevance and engagement. Analyzing the complete user journey on the platform will help us identify key touchpoints for effective ad insertions and improve overall user retention.

Finally, we will engage in regular benchmark studies against competitors of certain advertiser and keep abreast of market trends to adapt our strategies promptly, ensuring our methods remain innovative and aligned with industry shifts. Through these efforts, we aim to enhance the impact of our campaigns and drive better outcomes for both advertisers and users, keeping our strategies adaptable in the rapidly evolving digital advertising landscape.

Model Evaluation and Improvement

Firstly, while our model is capable of generating the best advertising strategies, verifying their applicability in real-world scenarios and the usability of the recommended parameters will

necessitate human evaluation. Collaborating with industry experts will be crucial to ensure that the strategies developed by our model are not only theoretically sound but also practically viable. For example, in a case study, our model recommended targeting devices like Connected_TV with a target age group of 13-26 for a mobile casino app. However, this target age is not legally permissible for gambling apps, and Connected_TV would not be an appropriate channel given the mobile nature of the app. This highlights the need for expert review to identify and rectify such mismatches, ensuring that the model's recommendations are both effective and compliant with industry standards and regulations.

Secondly, improvements in predictive modeling will involve adopting more sophisticated approaches. We plan to explore advanced predictive models such as neural networks and transformers, which can process complex data and patterns more effectively. Additionally, we aim to develop hybrid models that combine different types of modeling techniques, such as integrating machine learning models with rule-based systems. This combination will enhance the robustness and accuracy of our predictions, adapting more dynamically to diverse advertising challenges.

Thirdly, enhancements in deployment will focus on incorporating dynamic feedback mechanisms. Instead of relying solely on static predictions from exhaustive combinations, we intend to integrate deep reinforcement learning. This will allow our models to learn continuously from real-time data and adjust strategies dynamically, enhancing responsiveness and effectiveness in live environments. This approach will significantly refine how strategies are

tested and deployed, ensuring they evolve in response to changing market conditions and consumer behaviors.

10. Conclusion

In conclusion, our comprehensive exploration and implementation of predictive modeling for advertisement data prediction have demonstrated significant potential in optimizing online advertising strategies. The project, spearheaded by our dedicated team at Meetsocial, focused on developing a robust predictive model to forecast advertising outcomes for SMEs, particularly in app promotion. Throughout this journey, we delved deeply into various stages of data handling, from understanding and preparing the data to developing and evaluating the model.

The findings from our model development and evaluation phases underline the importance of choosing the right model and techniques based on the specific needs of the data and the business objectives. For instance, the use of XGBoost and Random Forest proved to be particularly effective in handling our skewed data distribution and complex feature interactions, leading to improved prediction accuracy as evidenced by our metrics, RMSE and R^2.

Additionally, our project has not only provided valuable insights into the factors driving advertising performance but also highlighted the critical role of data-driven decision-making in today's competitive business environment. By implementing these models, Meetsocial can offer SMEs more precise, cost-effective advertising solutions that are less reliant on trial-and-error methods and more grounded in analytical rigor.

Furthermore, the deployment plan ensures that these insights are translated into actionable strategies through a user-friendly dashboard, making the sophisticated models accessible to all stakeholders. This integration into the daily business processes at Meetsocial is expected to enhance operational efficiencies and client satisfaction.

Looking forward, there are numerous opportunities for further refining our models and expanding their capabilities. Continuous learning and adaptation, guided by new data and feedback, will be key to maintaining the relevance and effectiveness of our predictive models. By embracing these challenges, Meetsocial is well-positioned to maintain its leadership in providing innovative marketing solutions to Chinese businesses aiming to expand globally.

Appendix

Team Member's Contribution:

Team Member	Contribution
Leon Zhang	Dataset preparation and cleaning, Feature Engineering - building the scrapers for numerical features, Exploratory Data Analysis, Initial Modeling with Linear, Lasso and Ridge Regression. Modeling XGBoost model with new target Click/Impressions. Business understanding and data understanding (ppt and report).
Ishan Miglani	Dataset Merging, Dataset preparation and cleaning, Feature Engineering - building the scrapers for numerical features, Exploratory Data Analysis, Initial Modeling with Linear, Lasso and Ridge Regression.Helped with Modeling XGBoost model for all 5 targets and applying Hyper-parameter tuning. Data Preparation and EDA (ppt). Data Preparation and EDA(report). Created Ensemble models for all targets and extracted feature importances.
Keya Keya	Datasets merging, XGBoost Regression Model initial development, Hyperparameters Tuning, Model Overview (ppt), Reevaluation through Inverse Transformation,Model Development and Model Evaluation (report), Insights and Business Impact, Deployment Plan, Content outline and Final report formatting.
Qing Shen	Data collection, Dataset merging, Initial Modeling with Random Forest, Overfitting checking, Case study design and implementation (ppt and report), Future work analysis (ppt and report).

Link to Raw Data :
https://drive.google.com/drive/folders/1wMZ3WxQd_6otnf49zi18KBgfn5N627hW?usp=sharing
Link to the Code :
https://drive.google.com/drive/folders/1TZz1caiZOCX0knP0yZAUHD-1MX4Xe4tL?usp=sharing