Ad revenue prediction and Reserve price calculation

Problem Analysis

The problem revolves around optimizing revenue for a digital ad publisher operating in a real-time first-price auction environment. In such auctions, advertisers bid for ad slots, and the highest bidder wins and pays their bid amount. The key objectives are to estimate the potential revenue range for the upcoming july month and propose an optimal reserve price strategy based on past performance, bid shading patterns, and publisher revenue trends.To estimate the revenue range for July and determine optimal reserve prices using two different approaches: a time series forecasting method and a machine learning-based prediction model. The dataset includes revenue and ad impression data across various websites and ad types, allowing us to derive meaningful insights.

Data Exploration and Preprocessing

Before applying any predictive modeling techniques, the data was first checked for inconsistencies and missing values. Basic preprocessing steps included:

- Examined missing values and data distributions.
- Performed feature engineering to calculate CPM (Cost per Mille) as (total_revenue / total_impressions) * 1000 and handeled zero devision error.
- Created time-based features from Date column.
- Ensuring consistency in categorical variables.
- Performed EDA analysis.

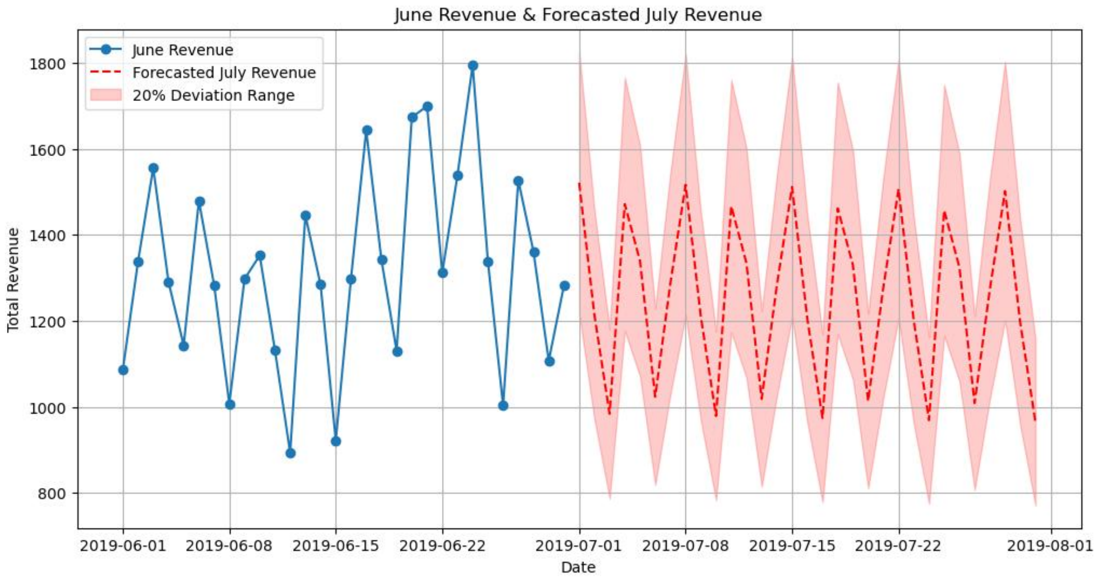
Time Series Analysis

Rationale: Since we have daily revenue data for June, time series methods can capture temporal patterns. The dataset was grouped by date, and the total revenue was summed over each day.

Stationarity Check: To determine whether the dataset is stationary, the Augmented Dickey-Fuller (ADF) test was performed. The data was found to be non-stationary,so first differencing was applied to make it stationary.

Model selction: Exponential Smoothing (ETS) method is used because it works well with limited data points, handles trend and seasonality components effectively and more interpretable than ARIMA for short series

Revenue prediction: Using the Exponential Smoothing model, total revenue for July was forecasted. The potential revenue range was calculated as **31,062$-46,593$** and a Seaborn (sns) plot was used to visualize the total revenue for June and the predicted revenue for July within a ±20% bands.

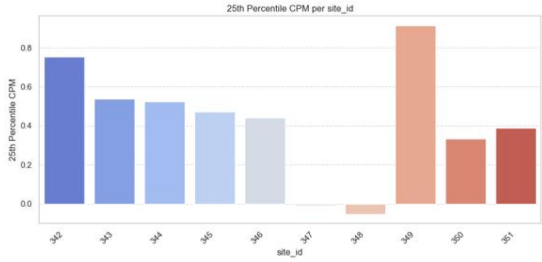


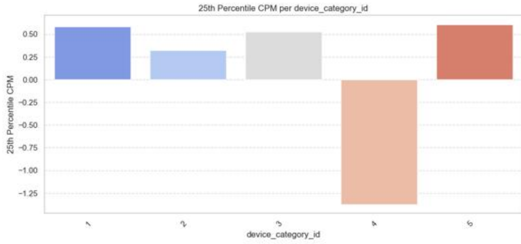
Machine Learning Analysis

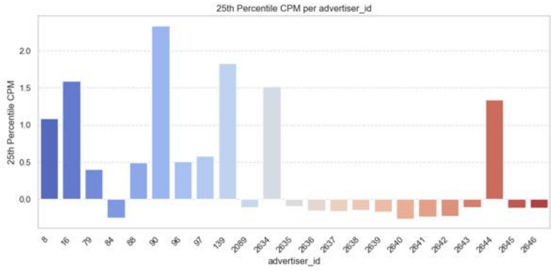
Rationale: Given the dataset's multidimensional nature (15+ feature dimensions) and limited temporal observations post-aggregation, time series models cannot inherently leverage the full feature set. Consequently, a machine learning approach was adopted. XGBoost was selected as the computational framework due to its demonstrated efficacy in handling mixed categorical and numerical features, its capacity to quantify feature importance for revenue impact analysis, robustness to outliers compared to linear models, and ability to model nonlinear feature interactions. Additionally, its integrated L1/L2 regularization mitigates overfitting risks, ensuring generalizability despite the high-dimensional input space.

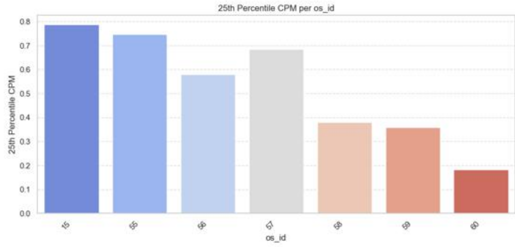
Revenue Prediction Using XGBoost: XGBoost was trained on key features with hyperparameter optimization performed via GridSearchCV to ensure model robustness. The model achieved strong predictive performance, with evaluation metrics of **MAE: 0.008** and **RMSE: 0.14**, indicating high accuracy. The trained model was then applied to predict July's total revenue, yielding a predicted revenue range of **31,630$- 47,445$**, derived from prediction intervals to account for uncertainty while maintaining the publisher's ±20% value deviation requirement. Actual revenue is 1% of the predicted revenue since Revenue_share is 1%.

Reserve Price Calculation: XGBoost was used to predict the reserve price using CPM as dependent variable. The dataset was grouped by individual key **c**ategories and detail statistics was evaluated for each category. The 25% quantile method was used to determine the lower bound for reserve price within each category. The 25th percentile CPM ensures that only the very lowest bids are filtered out, keeping the auction competitive while increasing revenue. We should follow the dynamic approch to set the reserve price. The calculated 25 percentile Reserve price in CPM for **site: 0.33-0.91, ad_type:0.5$, device_category: 0.32$-0.6$, advertiser: 0.5$-2.33$, os:0.18$-0.78$.**


25th Percentile CPM per site_id
25th Percentile CPM
0.0
0.2
0.4
0.6
0.8
342
343
344
345
346
347
348
349
350
351
site_id


25th Percentile CPM per device_category_id
25th Percentile CPM
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
1
2
3
4
5
device_category_id


25th Percentile CPM per advertiser_id
25th Percentile CPM
0.0
0.5
1.0
1.5
2.0
8
16
79
84
88
90
96
97
139
2089
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
advertiser_id


25th Percentile CPM per os_id
25th Percentile CPM
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
15
55
56
57
58
59
60
os_id