Evaluating Promotional Impact on Sales: A Case Study for Invent.ai's Data Scientist Role

Pouya Ghahramanian
ghahramanian@bilkent.edu.tr
Ankara, Turkey

ABSTRACT

This study investigates the impact of promotional campaigns on product and store sales, and develops a predictive framework to forecast future sales using real-world transactional data. In Part A, we analyze item- and store-level responsiveness to promotions by quantifying promotion lift and clustering entities into fast, medium, and slow movers. Our findings reveal substantial variation in promotion effectiveness across product groups and store locations, with certain categories showing strong positive responses and others demonstrating negligible or even negative effects.

In Part B, we build and evaluate machine learning models to forecast future sales over a four-month period using an eight-month training window. We enhance the dataset with temporal and promotional features and train LightGBM, XGBoost, and CatBoost regressors. Among these, the LightGBM model with full feature enrichment achieves the best performance, with an MAE of 1.61 and RMSE of 3.61. Results indicate the critical role of feature engineering and the presence of concept drift during promotional periods, highlighting the need for adaptive modeling strategies. All code and data are available at https://github.com/PouyaGhahramanian/promotion-impact-analysis.

ACM Reference Format:
Pouya Ghahramanian. 2025. Evaluating Promotional Impact on Sales: A Case Study for Invent.ai's Data Scientist Role. In *CIKM '25: ACM International Conference on Information and Knowledge Management, November 10–14, 2025, Seoul, Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 PART A: THE EFFECT OF PROMOTION ON PRODUCTS AND STORES

In this section we analyze the effect of promotion on products and stores. First, we describe our data preprocessing steps, then we answer the seven questions in this part (a-g) in the subsequent subsection.

1.1 Data Preprocessing and Steps

The date formats in the dataset are inconsistent. Specifically, the formats used for promotions 1–4 differ from those of promotions 5 and 6. I assume that the date format for Promo 5 is the same as that of Promo 6 (`dd-mm-yyyy`), since Promo 5 corresponds to the test set in the `1b.csv` file, which spans the 8th to 12th month of the year. This assumption ensures proper temporal alignment when analyzing the effect of promotions on the test set. Apart from this date format inconsistency, the dataset is well-structured and clean, and no additional preprocessing steps were necessary for the analysis.

1.2 Answers to the questions

In this subsection we provide answers to each of the questions in the part A of the assignment. Each question is given in bold text, followed by an answer to that question.

(a) What are your criteria for separating Fast, Medium, and Slow items? Why?
(b) What are your criteria for separating Fast, Medium, and Slow stores? Why?

To separate items and stores into Fast, Medium, and Slow groups, I used quantile-based clustering. Specifically, I computed the average weekly sales per store (for items) and per item (for stores) during non-promotion periods. Items (or stores) with averages below the 33rd percentile were classified as *Slow*, those between the 33rd and 66th percentiles as *Medium*, and those above the 66th percentile as *Fast*.

This approach is both simple and effective: using quantiles ensures robustness to outliers and avoids making assumptions about the underlying distribution of sales. Additionally, it provides a balanced grouping of the entities, which is especially useful for downstream modeling or evaluation.

Figure 1 visualizes the resulting clusters for both items and stores, showing their distribution and the threshold values used. A logarithmic scale is applied to the y-axis in both plots to reduce the visual distortion caused by extreme outliers in the Fast group and to make the cluster boundaries more interpretable.

(c) Which items experienced the biggest sale increase during promotions?
(d) Are there stores that have higher promotion reaction?

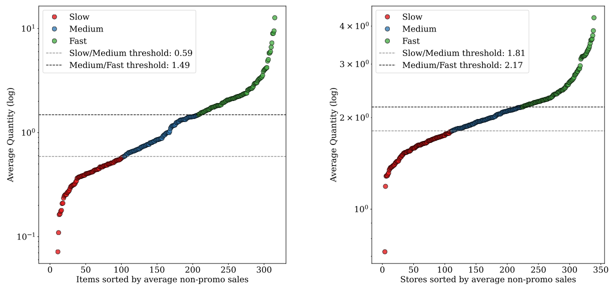
To quantify the impact of promotions, I calculated the *promotion lift* for each item and store as the difference between their average sales during promotion and non-promotion periods. Table 1 summarizes the top 10 items and stores with the highest promotion lift, along with overall statistics. These values reveal both highly responsive and negatively affected items and stores, highlighting the heterogeneous impact of promotions across the dataset.

Figure 2 illustrates the top 20 items and stores with the highest promotion lift, showing how sharply certain products and locations react to promotional campaigns.

To identify which types of products are more responsive to promotional activities, I analyzed the average promotion lift at both the

CIKM'25, November 10–14, 2025, Korea

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

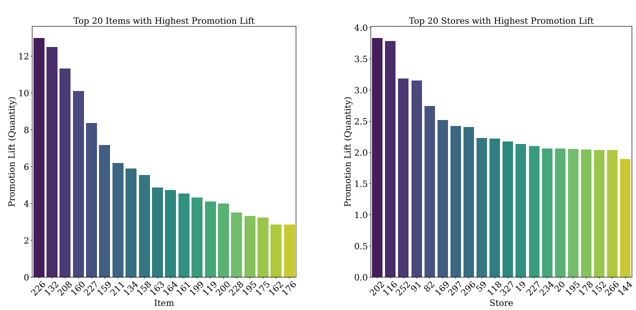


(a) Item clustering (b) Store clustering

Figure 1: Clustering of items and stores into Slow, Medium, and Fast groups based on average non-promotion sales. Dashed lines indicate the 33rd and 66th percentile thresholds. A logarithmic scale is used to mitigate the effect of outliers.

Table 1: Top 10 Items and Stores with Highest Promotion Lift, with Summary Statistics

Item ID	Promotion Lift	Store ID	Promotion Lift
226	12.97	202	3.83
132	12.48	116	3.78
208	11.33	252	3.18
160	10.11	91	3.15
227	8.38	82	2.74
159	7.18	169	2.52
211	6.19	297	2.42
134	5.90	296	2.40
158	5.54	59	2.23
163	4.86	118	2.22
Item Statistics		**Store Statistics**	
Min: -10.46	**Max:** 12.97	**Min:** -2.49	**Max:** 3.83
Mean: 0.29	**Std:** 2.35	**Mean:** 0.42	**Std:** 0.89



(a) Items (b) Stores

Figure 2: Top 20 items and stores with the highest promotion lift.

category level (ProductGroup1, PG1) and the subcategory level (ProductGroup2, PG2). As summarized in Table 2, Category A exhibits the highest average promotion lift (0.52), followed by Categories E and I. These results suggest that products in these groups tend to respond more positively to promotional campaigns.

At the subcategory level, PG2-5 stands out with the highest average lift of 0.64, followed by PG2-20 and PG2-4. In contrast, some subcategories such as PG2-18, PG2-1, and PG2-9 show negative lift values, indicating that promotions in these groups may lead to negligible or even reduced sales, potentially due to product returns or mismatched targeting.

These insights can be used to fine-tune promotion strategies by focusing on product segments with historically strong positive reactions, while exercising caution or reevaluating strategies for those with negative or minimal lift.

Table 2: Average Promotion Lift by PG1 Categories and Top/Bottom PG2 Subcategories

PG1	Lift	PG2 (Top)	Lift	PG2 (Low)	Lift
A	0.520	5	0.640	22	0.051
E	0.216	20	0.437	21	0.045
I	0.213	4	0.415	19	0.007
G	0.182	25	0.402	2	-0.012
J	0.118	26	0.241	11	-0.018
C	0.111	6	0.238	15	-0.027
D	0.073	14	0.234	7	-0.059
F	0.058	10	0.222	9	-0.064
H	-0.002	8	0.167	1	-0.085
B	-0.293	28	0.144	18	-1.334

Recommendation: The analysis reveals substantial variation in promotion effectiveness across items, stores, and product categories. Items in ProductGroup A and subcategories such as PG2-5, PG2-20, and PG2-4 show the strongest positive promotion lift, indicating a significant increase in sales during promotions. These items should be prioritized in future promotional campaigns to maximize impact.

Conversely, items in ProductGroup B and subcategory PG2-18 exhibit negative average promotion lift, meaning that returns during promotions exceeded additional sales. For these categories, promotions may not only be ineffective but also counterproductive. It is advisable to reassess or avoid promotional strategies for such items.

From a regional perspective, Stores 202, 116, and 252 demonstrate consistently high responsiveness to promotions and can be considered high-leverage targets for marketing investment. These insights support a more focused, data-driven promotion strategy aimed at products and locations with historically strong responses, thereby enhancing overall promotional efficiency and ROI.

(e) What is the biggest effect explaining sales change during promotions?

The analysis suggests that the biggest effect explaining sales change during promotions is the inherent responsiveness of specific items and stores to promotional activities. This is evidenced by the wide range of promotion lift values observed across items and regions, with some items (e.g., Item 226, Item 132) and stores (e.g., Store 202, Store 116) exhibiting significant sales increases, while others showed minimal or even negative responses. Additionally, item category plays a crucial role—certain categories (e.g., ProductGroup A and subcategories like PG2-5 and PG2-20) consistently demonstrate stronger lift, implying that product type is a key factor in promotional effectiveness. These variations highlight that promotion-driven sales changes are not uniform but depend heavily on product and location characteristics, making them the dominant explanatory factors for sales shifts during promotional periods.

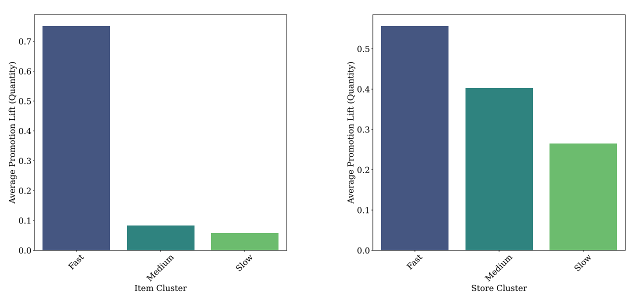
(f) Is there any significant difference between promotion impacts of the Fast versus Slow items?

(g) Is there any significant difference between promotion impacts of the Fast versus Slow stores?

To examine whether promotional campaigns affect item and store groups differently, I computed the average promotion lift across the three clusters (Fast, Medium, and Slow) for both items and stores. As shown in Figure 3, Fast items exhibited a substantially higher average lift of **0.75**, compared to only **0.08** and **0.06** for Medium and Slow items, respectively. This suggests that promotional activities are markedly more effective for fast-moving items.

A similar pattern is observed for store clusters. Fast stores experienced an average promotion lift of **0.56**, followed by Medium (**0.40**) and Slow (**0.27**) stores. While this trend is consistent, the magnitude of difference across item clusters is larger than that across store clusters.

Recommendation: These findings indicate a significant difference in promotional impact based on both item and store velocity. However, the item clustering explains a greater variation in promotional effectiveness, implying that *item dynamics are more predictive of promotion success than store performance*. Promotions should therefore be prioritized for fast-moving items to maximize return. The differences are illustrated in Figure 3, where a clear trend in promotional lift is visible across both dimensions.



(a) Average Lift by Item Cluster (b) Average Lift by Store Cluster

Figure 3: Average promotion lift by item and store cluster.

(Bonus) Is there any significant difference in item return rates after promotions?

To investigate whether item return rates change significantly after promotions, I analyzed the total return quantity as a function of the number of days since a promotion started (*PromoStartLag*). As shown in Figure 4, there is a striking peak on day 0, immediately after a promotion starts, with a total of 6,406 returns—substantially higher than the daily returns observed afterward, which drop below 500 from day 1 onward and steadily decline over the next 50 days.

This sharp spike suggests that a substantial volume of returned items is associated with immediate post-promotion sales, possibly due to over-purchasing, misaligned expectations, or stockpiling behavior during promotional periods. After day 0, the return rate gradually diminishes, indicating no sustained increase in return behavior post-promotion.

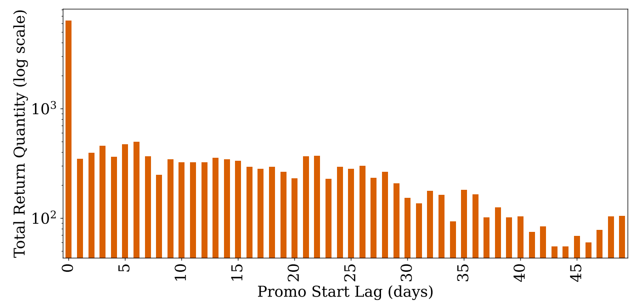


Figure 4: Total return quantity by days since promotion started (log scale).

Recommendation: Since a significant portion of returns occurs right after the promotion starts, retailers should consider implementing stricter return policies or more accurate promotion targeting to reduce immediate post-promotion returns. Additionally, monitoring return rates closely during the first few days of a campaign can help flag potential issues with specific items or stores.

2 PART B: MODEL TRAINING AND EVALUATION WITH PROMOTION DATA

In this section, I describe the process of training three regression models—LightGBM [3], XGBoost [1], and CATBoost [4]—on the training dataset *assignment4.1a.csv*, and evaluating their performance on the test dataset *assignment4.1b.csv*, which includes promotions 5 and 6. I begin by introducing a simple baseline model built upon the clustering approach developed in Part A. This baseline offers an interpretable point of comparison for the predictive models trained later in this section. Subsequently, I outline the feature enhancement steps and present the training and evaluation results of the predictive models.

2.1 A Baseline Model Based on Clustering in Part A

As a starting point, I develop a naive forecasting model using the item and store clusters created in Part A. The primary idea is to predict sales during Promotion 5 by leveraging the historical average sales of items in each cluster during non-promotional periods, and adjusting these with the corresponding promotion lift values computed per cluster.

To apply this model, each item and store in the test set is assigned to its cluster based on the mappings learned from the training data. For each item cluster, I calculate its average sales in non-promotion periods and add the average promotion lift that cluster historically experienced. This yields the expected sales quantity for Promotion 5:

$$\text{ExpectedQuantity}_{i,t} = \mu^{\text{non-promo}}_{\text{cluster}(i)} + \text{Lift}_{\text{cluster}(i)} \tag{1}$$

This approach assumes that past average performance and lift patterns are indicative of future behavior under similar promotional settings. While it does not take into account finer-grained patterns or covariates such as day-of-week effects or promotional timing, it provides a meaningful baseline. It allows us to evaluate whether more complex machine learning models can yield significant improvements over this interpretable and easy-to-implement benchmark.

Results. Table 3 summarizes the forecasting performance of the cluster-based baseline model. On Promotion 5, the model achieves a mean absolute error (MAE) of 2.72 and root mean squared error (RMSE) of 5.70. When evaluated over the entire test set (including both promotional and non-promotional periods), the model yields an MAE of 2.38 and an RMSE of 4.58.

Table 3: Baseline Model Performance

Metric	MAE (Promo5)	RMSE (Promo5)	MAE (All)	RMSE (All)
Values	2.72	5.70	2.38	4.58

Reference Scale: The average true sales value across the test set is 2.72, with a standard deviation of 5.93, and a range spanning from -12.00 (indicating returns) to a maximum of 192.00 units sold. This highlights the high variance and skewness in sales data, making it challenging to model precisely. Nevertheless, the baseline model performs reasonably well in this context, and sets a performance bar for the more advanced models discussed in the next sections.

2.2 Feature Engineering

To improve model performance and better capture sales patterns, I applied several feature engineering steps to enrich the data with informative variables. These features aim to capture seasonality, temporal trends, store-item dynamics, and promotion timing effects.

The following features were added:

- **DayOfWeek**: The day of the week (0 for Monday, 6 for Sunday) extracted from the date. Captures weekly seasonality.
- **isWeekend**: A binary feature indicating whether the date is a weekend (Saturday or Sunday).
- **Last7Avg**: Rolling 7-day average of sales quantity per store-item pair. Reflects short-term trends.
- **Last30Avg**: Rolling 30-day average of sales quantity per store-item pair. Captures medium-term patterns.
- **LastSaleDayDiff**: The number of days since the last recorded sale for each store-item pair. Highlights sparsity and item activity.
- **PromoStartLag**: Number of days since the last promotion began for the given store-item pair. Derived by tracking when promotions start and forward-filling the most recent start date.
- **StoreCluster** and **ItemCluster**: Cluster assignments obtained from Part A based on average sales quantities, used as categorical indicators.

For the rolling average features (`Last7Avg`, `Last30Avg`), I assumed a causal structure where daily statistics are available only after the given day, ensuring realistic test-time availability.

Figure 5 displays the feature importances derived from a LightGBM model trained on the processed dataset. As seen, `Last7Avg` is by far the most influential feature, followed by `LastSaleDayDiff`, `DayOfWeek`, and `PromoStartLag`, all of which capture temporal or promotional effects. Interestingly, the cluster indicators (`ItemCluster`, `StoreCluster`) were found to have relatively low predictive power compared to the time-series derived features.

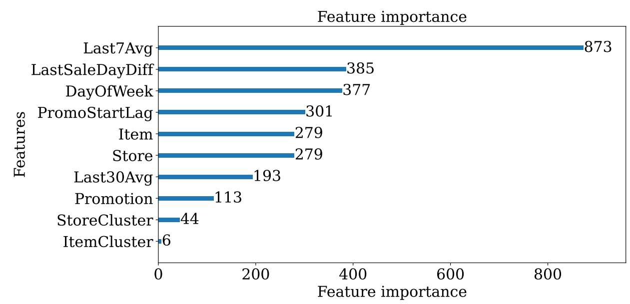


Figure 5: Feature importances from the LightGBM model

2.3 Experimental Results

In this section, we evaluate the performance of three popular gradient boosting models: **LightGBM** [3], **XGBoost** [1], and **CatBoost** [4]. Each model is trained on the training set and evaluated on the test set using three different feature configurations:

1. **All Features (AF):** The complete enhanced feature set, including temporal averages (Last7Avg and Last30Avg), item/store clusters, and promotion flags.
2. **No Temporal Averages (NoAvg):** The enhanced feature set excluding Last7Avg and Last30Avg, assuming that recent statistics may not be available for prediction at each day.
3. **Original Features (OrgFeat):** The original, unenhanced feature set including only Item ID, Store ID, Item Cluster, Store Cluster, and Promotion indicator.

A validation split of 10% from the training set was used during training. Table 4 presents the results of each model under the different feature settings. Among them, the **LightGBM model with the full feature set (AF)** achieves the best performance across all metrics. Notably, all models show substantial improvement when the full feature set is used.

For example, the LightGBM model with AF achieves approximately **20%** lower RMSE compared to the NoAvg setting, and **21%** lower RMSE compared to the OrgFeat setting. This indicates the predictive value of incorporating recent average sales statistics, as they help the model learn seasonal patterns and short-term trends. The importance of these features was also highlighted in Figure 5.

Higher MAE and RMSE values observed specifically during the Promo5 period—compared to the full test set—suggest that this period differs from the rest. This supports prior findings such as those in [2], indicating the presence of *concept drift* due to promotion

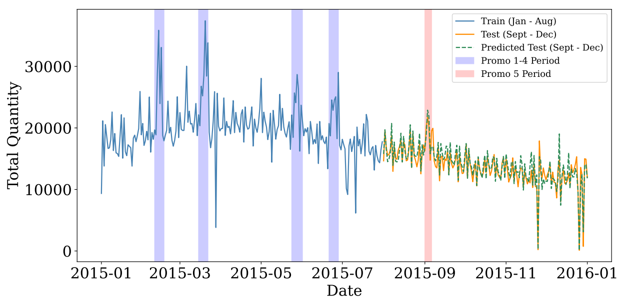
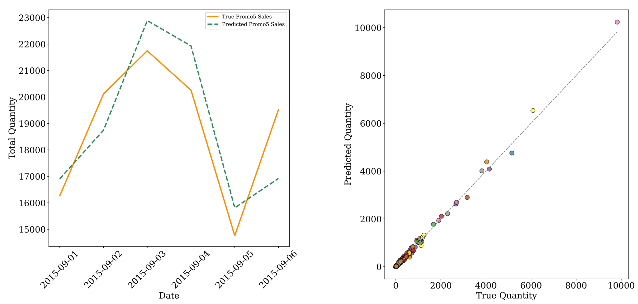


Figure 6: True vs. Predicted total sales during the test period using the best model (LGBMR with AF).



(a) Total quantity over time

(b) Quantity per item

Figure 7: True vs. Predicted values using LGBMR (AF) during the Promo5 period.

effects. This emphasizes the need for continual model adaptation to remain accurate in the presence of distributional changes.

Reference scale of true target values: mean = 2.72, std = 5.93, min = −12.0, max = 192.0.

Figure 6 shows the predicted vs. actual total sales over the test period (including the Promo5 period), using the best-performing model configuration. The model closely tracks the true values, confirming its predictive strength even during promotion periods.

Figure 7(a) illustrates the total sales predictions over time during the Promo5 period, while Figure 7(b) compares predictions at the item level. Most predictions are tightly clustered around the $x = y$ line, indicating accurate item-level forecasting. All results are based on the LightGBM model with the complete feature set.

Goodness of Fit Measures: We use **MAE** and **RMSE** to evaluate model performance. For example, an MAE of 1.61 means that the model's predictions deviate from the true sales by an average of 1.61 units.

How good is the model from Step 1? The baseline model from Step 1 is included as the first row in Table 4. Models in Step 2 clearly outperform the baseline due to the use of enhanced features and more powerful learning algorithms.

Main Problem Points: One major issue is the *distributional shift* over time. For instance, the impact of Promo5 may differ from earlier promotions due to seasonal changes (e.g., spring vs. autumn) or macroeconomic factors affecting consumer behavior. If historical data from previous years were available, models could better capture such recurring patterns.

What would you change in Step 1? The naive baseline model served as a reference. However, to improve it, one could incorporate additional features or use more robust modeling techniques, even at the initial stage.

Is there any data set that you would like to use in addition to tables provided for this assignment? What are those data sets? How would you obtain them?

Yes. I would consider:

- **Calendar and Economic Indicators:** Public holidays, special events, and inflation data from official government sources or open APIs.
- **Historical Sales Information:** Sales data from previous years for the same periods, ideally provided by the company.
- **Promotion Information:** Detailed metadata about past promotions (e.g., discount rate, marketing channel), obtained from internal marketing records.

Table 4: Results of all models under different feature settings. Best results for each metric are in bold.

Model	MAE (Promo5)	RMSE (Promo5)	MAE (All)	RMSE (All)
Baseline	2.72	5.70	2.38	4.58
LGBMR (AF)	**1.96**	**4.39**	**1.60**	**3.61**
LGBMR (NoAvg)	2.45	5.29	1.96	4.21
LGBMR (OrgFeat)	2.48	5.32	2.02	4.26
XGBR (AF)	1.97	4.47	**1.60**	3.64
XGBR (NoAvg)	2.46	5.27	1.95	4.20
XGBR (OrgFeat)	2.48	5.30	2.02	4.26
CBR (AF)	1.98	4.45	1.62	3.65
CBR (NoAvg)	2.48	5.33	1.99	4.26
CBR (OrgFeat)	2.50	5.35	2.04	4.29

3 CONCLUSION

In this work, we developed a predictive model to forecast future sales of a company over a four-month test period using eight months of historical training data. As a preprocessing step, we clustered both items and stores into fast, medium, and slow categories based on their historical sales statistics, and assigned corresponding cluster labels to the test data. Next, we applied a feature enhancement procedure to enrich the feature set with temporal and contextual features to improve model performance. We then trained three gradient boosting models—LightGBM, XGBoost, and CatBoost—on the enhanced feature set.

Our best model achieved a Mean Absolute Error (MAE) of 1.61 and Root Mean Square Error (RMSE) of 3.61 on the full test set, and 1.96 and 4.39 respectively on the Promo5 period. These results demonstrate the effectiveness of our feature engineering strategy and the utility of gradient boosting models for sales forecasting under promotion-induced variability.

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[2] Pouya Ghahramanian, Sepehr Bakhshi, Hamed Bonab, and Fazli Can. 2024. A novel neural ensemble architecture for on-the-fly classification of evolving text streams. *ACM Transactions on Knowledge Discovery from Data* 18, 4 (2024), 1–24.

[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[4] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).