Spring 2023 Capstone Project Final Report

DSI x L'Oréal: Time Series Sales Prediction for Kiehl's

Columbia University

Yifan Lu (yl5113), Jianing Yu (jy3266), Ruijie Zhang (rz2596),

Yancheng Zhang (xz3157), Yifan Zhu (yz4360)

May 12, 2023

1 Introduction

1.1 Background and Motivation

The time series prediction helps better decision making on overall project planning, logistics, inventory, and risk management. Sales forecasting allows companies to efficiently allocate resources for future growth and manage its cash flow, which is especially important for industries like cosmetics. Significant surges in sales data are noticeable during pre-holiday, and some seasonal trends are captured which present a challenge for any traditional forecasting models since observations are dependent.

Kiehl's was acquired by L'Oréal in 2000, since then it skyrocketed from a reported sales figure of €60 million to €960 million, boosted notably by the opening of mono brand stores. This capstone is sponsored by L'Oréal, with the motivation to predict sales volume for L'Oréal's subsidiary brand Kiehl's, which could help to gain insights into customer behavior to optimize future business decisions.

1.2 Research Question

Our initial goal is to construct a daily sales prediction model for 68 stores across the United States for Kiehl's, with customized time windows ranging from 1 to 365 days ahead on a store-by-store basis. Due to delay in store-level sales data acquisition, we currently emphasize on traffic prediction after verifying high correlation between sales and traffic.

1.3 Literature Review

Time-Series Forecasting is an important topic in data science. Existing work on time-series sales forecasting generally focuses on these models:

1. The LightGBM model is an ensemble learning model based on Gradient Facilitated Decision Tree (GBDT). The LightGBM model also uses the idea of mutually exclusive feature bundling (EFB), which combines some mutually exclusive features to reduce the dimension of features. Finally, the LightGBM model adopts the leaf-wise leaf growth strategy with depth limitation [1]. For example, Zhang et al. used the LightGBM model to mine the nonlinear relationship in sales and combine it with LSTM to predict supermarket vegetable short-term sales [2]. Generally, Light GBM outperformed many other ensemble models because of its high efficiency, but it is noticeable that light GBM is particularly sensitive to overfitting because of the complexity of the trees [3].
2. Autoregressive Integrated Moving Average (ARIMA): Ramos et al. [4] used ARIMA and state-space models to predict the future sales volume of commodities. ARIMA performs well on time series that typically contain trend and seasonal patterns. However, ARIMA models cannot capture the interactions and dependencies between different variables, such as the effects of external factors on the time series.

Other than these previous studies, there are widespread applications of machine learning in the prediction of time series problems, for example, Vallés-Pérez el at. [5] proposed a prediction based on recurrent neural networks to predict large grocery sales. Based on our reading of

previous research, many different models have been implemented for sales forecasting problems. Because these models have unique strengths and weaknesses, they behave differently for different application scenarios and data. Therefore, we will firstly explore the trends and characteristics of our dataset, then implement multiple models mentioned above, including but not limited to LightGBM, ARIMA, neural networks, to investigate which model is the best fit for the L'Oreal dataset.

1.4 Overall Approach

Since a strong positive correlation between traffic and sales was captured, the overall approach could be broken down into two stages: traffic prediction and sales prediction. Firstly, a light gradient-boosting machine (Light GBM) was proposed among several alternative models to predict traffic. Then, we could either feed this prediction as an input variable to the following sales forecasting model or adopt similar architecture on sales prediction. The performance of these models was compared with a naive baseline model based on several defined evaluation matrices.

2. Methodology

2.1 Dataset and Exploratory Data Analysis

In this project, we utilize 3 datasets: country-level sellout data, country-level, traffic data, and state-level traffic data within the United States. The main objective of analyzing country-level sellout data and traffic data is to explore the correlation between number of visitors and actual sellout amount. This analysis helps to identify trends and patterns that can influence the overall performance of the business.

On the other hand, the state-level traffic data is merged with census data and external data sources to create a more comprehensive dataset for traffic prediction. By integrating external data sources, accounting for regional variations, and incorporating additional features such as

holiday encoding and time transformations, we strive to enhance the accuracy and robustness of our predictive models. This approach ultimately leads to a better-informed decision-making process and optimized resource allocation, facilitating the achievements of Kiehl's business goals.

We will provide a detailed overview of the three datasets based on their specific applications in the following sections:

2.1.1 Datasets for Detecting Correlation Between Sales and Traffic

Due to existing accessibility, we first obtained sales data and traffic data at country level (for 14 countries), which serves as a reference for our research. The traffic dataset ranges from 2017-Dec-31 to 2022-Nov-03 containing 14834 observations, while the sales dataset ranges from 2017-Jan-01 to 2021-Dec-31 containing 19071 observations. Although the dataset is not suitable for predicting store-level data, it remains a valuable resource for correlation investigation.

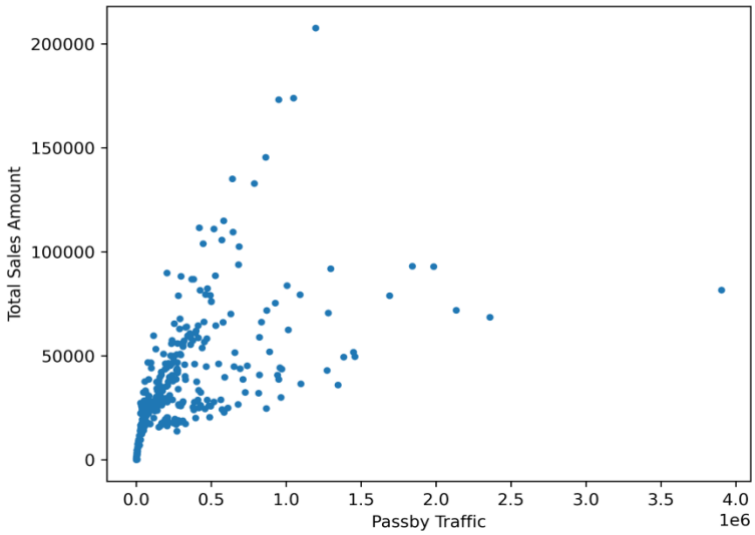

Fig1: Sellout vs. Passby Traffic

A basic scatter plot illustrating the relationship between sellout and pass-by traffic can be useful in establishing correlation. While the connection may not be strictly linear due to the

time series nature of the data, this visualization can still serve as a useful foundation for forecasting sales when employing more advanced techniques.

To provide a more comprehensive overview of the trend, we conducted a comparative analysis of sales and traffic data using a time series plot. The plot reveals a strong positive correlation between the sellout amount and passby traffic. We further annotated the graph with holidays, including New Year's Day, Valentine's Day, Black Friday, Thanksgiving Day, Christmas Day, and the two days before and after these holidays, which are represented by gray shaded areas.

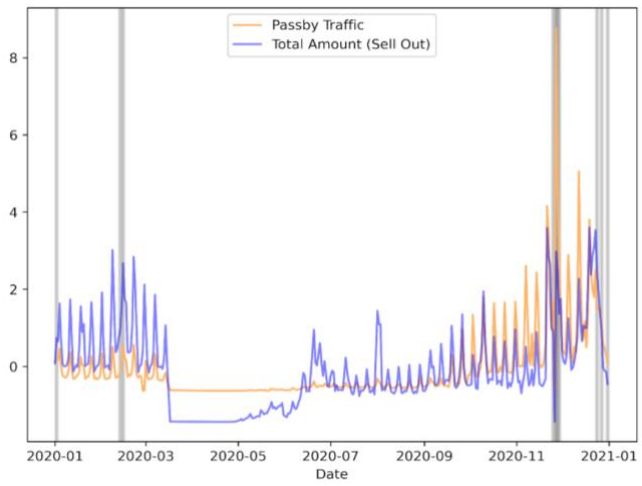

Fig2: Sellout vs. Passby Traffic Over Time

From the observation, the day prior to each holiday displays a peak in both the sellout amount and passby traffic, as indicated within the grey shaded regions. This observation underscores the necessity of incorporating holiday-related encoding features in our dataset to accurately capture the observed trend and enhance the predictive power of our model.

2.2.2 Dataset used to predict traffic

The dataset used in this stage consists of two primary parts. The first part contains sales data provided by L'Oreal, spanning from 2015-Apr-21 to 2019-Aug-18. The dataset contains 188951 rows and 14 columns. To simulate a real-life business scenario, we utilized data from 2016 to 2018 as the training set and 2019 data as the test set. Some columns have been

transformed to enhance model accuracy. The second part of the dataset includes external census data and holiday indicators, which have been integrated based on city and state.

Following data processing and filtering for the Kiehl's brand, the dataset comprises 87,126 rows and 184 columns. The large number of columns is primarily attributable to the one-hot encoding applied to categorical variables. Notably, there are no missing values in the modeling section.

2.2 Data Pre-processing

Data representative: Though the datasets do not directly contain null values or outliers, there are many 0s that possibly represent missing values; thus, we filled those cells with mean value before training the model. Also, after investigating different time ranges for each store, we filtered data after '2016-01-01' since data in 2015 only represents the information at desert hill outlets in California. Potential relevant variables are selected with preprocessing steps below.

Geographic Variables: To further improve the performance of predicting on unseen data, variables 'state' and 'city' were encoded by external census data. More specifically, we used the population size and household income level of each state and city to create a set of features that can be used as input for our model.

Sine and Cosine Transformation: Since time series data type is not independent, sine and cosine transformations can convert 1D datetime into 2D arrays without losing this information. The day of the year could be decomposed into a sine wave and a cosine wave, with the amplitude and phase of each wave representing cyclic patterns of the year.

Label Holidays: We used holiday calendars and traffic data visualization results to identify important holiday periods that may affect store traffic. We observed there are three significant peaks that are not necessary on holidays, so we needed to manually label them: the day after Thanksgiving (Nov.23 – Nov.25), Christmas (Dec.22 – Dec.23), the week before Christmas (Dec.16 – Dec.17).

Lookback period and rolling mean window: We self-defined function to add the value from previous timestamp as new columns for each observational day. Those number of lookback days [1, 4, 5, 6, 7, 12, 14, 21, 27, 28] were tuned and selected based on feature importance.

2.3 Analytics Approach

There are multiple stages of the model development. Firstly, we observed the significant correlated trend between total sellout and traffic from visualizations and correlation test; thus, our initial goal for the first stage is to construct a traffic prediction model based on time series data. Then, the estimated traffic will be fed forward to the next sellout model, which will be developed in the second half of the semester.

The methodology implemented herein are constructing a naïve baseline model using historical average from previous years, and using static regression model (generalized linear model, GLMM), time series model (Autoregressive integrated moving average, ARIMA), and advanced machine learning models (e.g., Light GBM, XGBoost, LSTM, Prophet) to further improve the evaluation matrix.

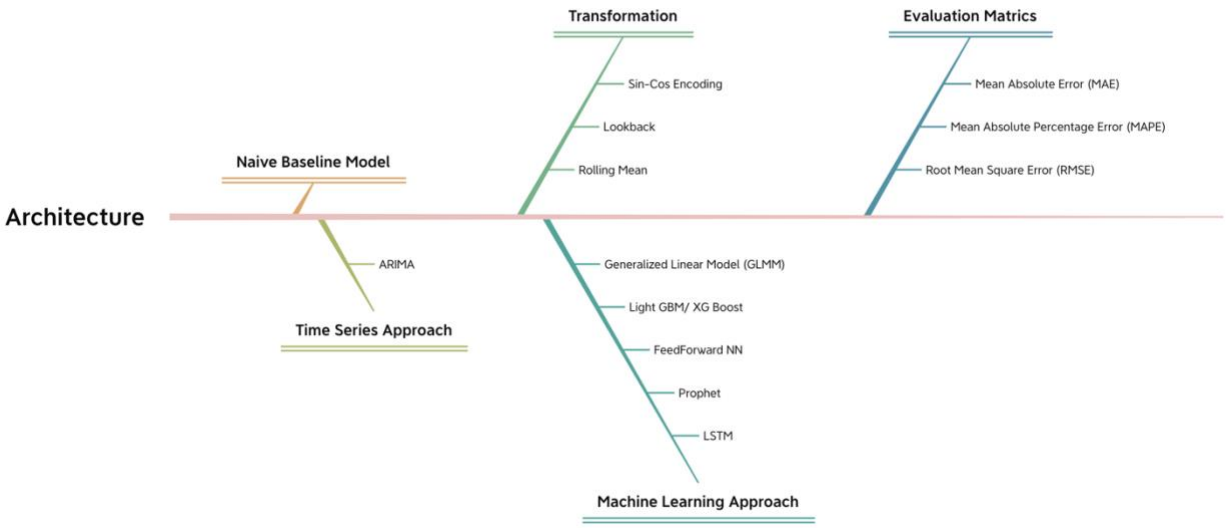

Fig. 3 Summary of Overall Approach

2.3.1 Baseline Model

The initial baseline model predicts traffic by taking the average of traffic in historical data on a given date. This approach is considered naive since it assumes no major changes or external factors affecting the store's traffic patterns, which is not practical in reality but helpful as a reference for further model construction and a benchmark to assess the efficacy of advanced models.

2.3.2 Light GBM

The algorithm uses a tree-based learning algorithm that leads to more efficient use of resources and faster training. To build our Light GBM model, we first preprocessed the dataset into 184 columns in total to incorporate relevant features and transformations, then divided the dataset into training (70%), validation (20%), and testing (10%) sets. L1 and L2 regularization was used to prevent overfitting. The final tuned model uses the following hyperparameters:

Objective: Regression
Metric: RMSE (Root Mean Squared Error)
Number of leaves: 31
Max depth: 8
Learning rate: 0.055
Number of estimators: 1000
L1 regularization term α: 0.4
L2 regularization term λ: 0.2
Number of estimators: 1000

Since we incorporate a lookback variable into the dataset, it is important to note that in real-world scenarios, obtaining such historical information becomes challenging when the target value is significantly distant from the present. To predict values that are far away from the current time, a recursive algorithm is necessary, which involves predicting earlier values before estimating the more distant ones.

2.3.3 Alternative Model

We also explored some complex models to predict future store traffic, such as Generalized Linear Mixed Models (GLMM), ARIMA, and Prophet so far. But the results are worse than the LightGBM model. Here are some brief summaries of the models:

GLMM:

$$y_{ij} = \beta_0 + \beta_1 State_i + \beta_2 Flasghip_i + \beta_3 SalesZone_i + \beta_4 Holiday_i + \beta_5 weekday_i + \beta_6 month_i + \beta_7 Population_i + \beta_8 MHI_i + u_j + \epsilon_{ij}$$

ARIMA:

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \cdots)(1 - B) y_t = (1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3 + \cdots)\epsilon_t$$

where:
B is the backshift operator, where $By_t = y_{t-1}$
$\phi_1, \cdots, \phi_5$ are the AR coefficients
$\theta_1, \cdots, \theta_5$ are the MA coefficients
ϵ_t is white noise with zero mean and constant variance

PROPHET:

We use Prophet to predict traffic based on a set of predictor variables, including the year, month, population, MHI, and several variables related to holidays and weekdays.

2.4 Evaluation Method

We employ evaluation metrics such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to assess the performance of our prediction models. These metrics are essential for gauging the accuracy and effectiveness of the models, as they provide different perspectives on the errors:

MAPE measures the average percentage error between predicted and actual values, making it easy to interpret and allowing for comparison across different scales. MAE represents the average magnitude of errors in predictions, regardless of direction, offering an understanding of the overall prediction error. RMSE is the square root of the average squared differences between predicted and actual values, emphasizing larger errors and providing a sense of how significantly the model deviates from the true values.

By using these metrics together, we can obtain a comprehensive evaluation of our models' performance and make informed decisions on how to improve them.

Table 1: Model Evaluation

Model Comparison - Mean Absolute Percentage Error (MAPE) - Mean Absolute Error (MAE) - Root Mean Square Error (RMSE)	MAPE	MAE	RMSE
Light GBM - SinCos transformation on 365 days - Lookback on days - Rolling window mean	19.82	5.13	8.19
Light GBM	20.36	5.30	8.43
XGBoost			10.42
FeedForward NN - SinCos encoding			12.82
Prophet - With holiday parameter			22.09
Generalized Linear Model (GLMM)			23.93
Naïve Baseline - Fitting on previous year			25

2.5 Results

Overall, by using GLMM, Light GBM, and Prophet models, we can significantly improve the accuracy of our predictions compared to the baseline model, especially after adding a lookback period to the model. The current winner among all proposed models is LightGBM (with datetime sin-cos transformation, lookback and rolling mean window).

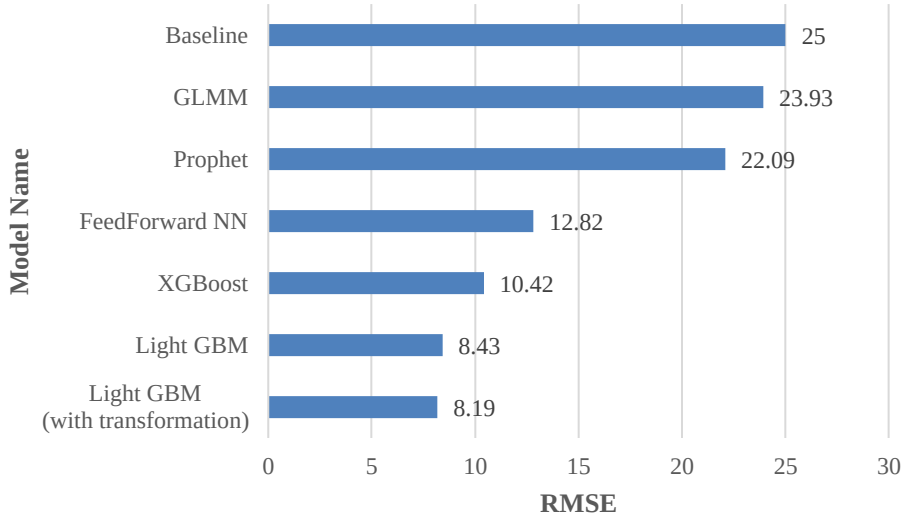

Fig4: Model Performance Evaluation

3. Discussion

During the latter half of the semester, our goal is to enhance the existing model's performance by incorporating potential external factors and optimizing hyperparameters. Subsequently, we will focus on creating an appropriate model for predicting sellouts, utilizing estimated traffic data. There are two potential approaches. First, we could consider predicted traffic as new input variable and feed it to sales model; or we could apply similar architecture on sales data since traffic and sales are highly correlated. However, these two hypotheses need to be verified in the future work.

4. Team Member and Contribution

- Yifan Lu (yl5113@columbia.edu): prepared software structure, wrote python scripts for backtesting internal macroeconomic index with external dataset
- Jianing Yu (jy3266@columbia.edu): main contributor to literature review, defined business problems and helped to aggregate available external data sources
- Ruijie Zhang (rz2596@columbia.edu): communicated with industry mentor, applied transformation on dataset, and constructed baseline model
- Yancheng Zhang (xz3157@columbia.edu): data visualization and data analysis in python, contributed to model selections
- Yifan Zhu (yz4360@columbia.edu): team captain, monitoring overall project logistics and progress; applied data preprocessing and developed advanced said models

References

[1] Tingyan Deng, Yu Zhao, Shunxian Wang, and Hongjun Yu. Sales Forecasting Based on LightGBM. 2021. https://ieeexplore.ieee.org/document/9342445/references#references

[2] Zhang He, Sun Yu. Application of LightGBM and LSTM combined model in vegetable sales forecast. 2020. https://iopscience.iop.org/article/10.1088/1742-6596/1693/1/012110/pdf

[3] Youyang Zhang, Changfeng Zhu, Qingrong Wang. LightGBM-based model for metro passenger volume forecasting. 2020. https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/iet-its.2020.0396

[4] Ramos P, Santos N, Rui R. 2015. Performance of state space and ARIMA models for consumer retail sales forecasting[J]. Robotics and Computer-Integrated Manufacturing,2015,34:151-163. https://www.sciencedirect.com/science/article/abs/pii/S0736584515000137

[5] Iván Vallés-Pérez, Emilio Soria-Olivas, Marcelino Martínez-Sober, Antonio J. Serrano-López, Juan Gómez-Sanchís, Fernando Mateo. Approaching sales forecasting using recurrent neural networks and transformers. 2022. https://www.sciencedirect.com/science/article/pii/S0957417422004146