Kamella Hoang

**Local Restaurant Sales Analytics**

December 5th, 2024

**I. Project proposal**

1. *Personal objective*

My primary objective is to get used to working with data in Tableau and improve my data visualization, data preprocessing, and predictive modeling skills. The project goal is to predict future sales trends for a local restaurant sales data in India and help this restaurant make great decisions to boost its sales and operate more effectively. Within the project, I will analyze restaurant data in 2 years and create predictive models predicting customer behaviors, and sales trends in the future.

1. *Intended outcomes*

+ **A high-accuracy predictive model**: a model that can predict future sales by multiple factors based on past data. In addition, the model can also predict what items will be popular or what time in the day or what period of the year the restaurant will be busy. This model will help the restaurant quickly adapt to different customer behaviors and sales trends.

+ **Key data insights**: find out the trends in customer behavior, such as which are the most popular items at different times or whether payment methods can affect sales. By doing so, the owner can implement proper marketing strategies and promotions.

+ **A** **dashboard:** use Tableau to create an interactive dashboard that shows both the sale trends and predictions, where the owner or the manager can access the data and assess what to do.

1. *Description of the needs of the intended audience*

The intended audience in this project is the restaurant owner and the restaurant managers. This project will help them make informed decisions on managing team members, menu optimizing, and marketing strategies:

+ **Prepare for peak hours, and peak times**: The predictive model will let them know what time of the day or what time of the year the sales will spike so the the manager can allocate employees effectively. This is important because if there are not enough employees working, the restaurant can lose lots of customers because they have to wait too long.

+ **Optimize the menu:** the predictive model as well as the key insights will predict which dish will sell well. In addition, knowing which dishes are unpopular can also help the restaurant find out why the dishes are not favored. They can make new dishes that are quite the same as the popular ones, and adjust the favors in the unpopular, or even remove them from the menu.

+ **Increase sales at different times**: Since the predictive model can predict the revenue, the owner can offer special promotions during slower times. In addition, the owner can also put the less popular dishes on sale.

1. *Foreseeable challenges*

+The predictive models need high accuracy of the data. If the data quality is bad, or it has a lot of missing values or inaccurate data, it is hard to clean the data. In the case that I cannot handle the data cleaning and data processing well, this can affect the prediction results.  
+ The data is collected within around 2 years, so the prediction can only be based solely on the 2-year data, which means data on some special occasions that often occur 4 years a time, or 2 years a time (such as World Cup or other champions) could be missed. As a result, when these events occur in the future, the predictions can be low in accuracy.

**II. About the dataset**

Taken in Kaggle: <https://www.kaggle.com/datasets/rajatsurana979/fast-food-sales-report>

This dataset of 1000 rows captures sales transactions from a local restaurant in India from April 2022 to March 2023. It includes details about the order ID, date of the transaction, item names (representing various food and beverage items), item types (categorized as Fast-food or Beverages), item prices, quantities ordered, transaction amounts, transaction types (cash, online, or others), the gender of the staff member who received the order, and the time of the sale (Morning, Evening, Afternoon, Night, Midnight). The dataset offers a valuable snapshot of the restaurant's daily operations and customer behavior.

* **Columns**:

1. *order\_id*: a unique identifier for each order.
2. *date*: date of the transaction.
3. *item\_name*: name of the food.
4. *item\_type*: category of item (Fastfood or Beverages).
5. *item\_price*: price of the item for 1 quantity.
6. *quantity*: how much quantity the customer orders.
7. *transaction\_amount*: the total amount paid by customers.
8. *transaction\_type*: payment method (cash, online, others).
9. *received\_by*: gender of the person handling the transaction.
10. *time\_of\_sale*: different times of the day (Morning, Evening, Afternoon, Night, Midnight).

**III. Prepare the dataset**

transaction\_type has missing values, all other attributes has 1000 rows. As the transaction\_type only has cash, and online, I will fill the missing values with the data in front of it, and behind it. I also changed it to proper data type, as date was not in proper format (shown in Python notebook)

1. **Interpretation of dataset skewness:** using only numerical features

- item\_price has a skewness of 0.624, since it is positive -> the distribution has a longer tail on the right side. It is slightly right-skewed but close to symmetric

- quantity has a skewness of -0.051, which is a slight negative skewness -> Distribution is quite symmetric, as the value is very close to 0

- transaction\_amount has a skewness of 1.051, which is positive -> Distribution is moderately reight-skewed, having a longer tail on the right side

**b) Normality visualization:**

A diagram and diagram of a graph

Description automatically generated with medium confidence A graph and diagram of a graph

Description automatically generated with medium confidence A graph and diagram of a graph

Description automatically generated with medium confidence

- item\_price: the distribution on histogram is right\_skewed, as indicated by a peak on the left (around 20) and a tail stretching towards higher values (60+). The KDE line shows that item\_price has non-normal distribution with multiples peaks. Based on the QQ plot, item\_price doesn't have normal distribution as it has right skewness

- quantity: For histogram, the distribution is quite uniform, with no clear peak or bellshape structure. The KDA line doesn't follow a normal bell curve. Based on QQ plot, the points generally align with the diagonal red line but show some deviation. The data is roughly symmetric and follows a normal distribution. Quantity is more symmetric than item\_price but lacks a clear normal distribution because of its uniform nature.

- transaction\_amount: Using the histogram, the distribution iss strongly right-skewed, with a peak at lower values and a long tail towards higher values. The KDE line shows the skewness and lack of symmetry. Based on the QQ plot, the points deviate significantly from the diagonal red line, particularly at the upper tail (right skew). It means transaction\_amount

**c) Transform the data for machine learning model**

A diagram and a diagram

Description automatically generated with medium confidence A graph and diagram of a graph

Description automatically generated with medium confidence A graph of a number of values

Description automatically generated with medium confidence

With the transformed data, it can be clearly seen the data is distributed equally, especially with the transaction\_amount. We will use this transformed transaction\_amount for machine learning.

**IV. Dashboard**

Using the data from the regular data to capture the real data

1. KPI Sales metric:

The restaurant achieved sales of $25,390 in March 2023.

Sales have decreased by 0.80% compared to the previous month, minor decrease

In April 2022: Sales began at 17,670 INR, indicating a lower point in the time frame. Despite some dips, sales have consistently grown over time, with $25,390 in March 2023, reflecting a strong upward trajectory overall. This is a positive sign that indicates the restaurant has likely expanded its customer base or improved its offerings.

2. Sales Performance over time and 5 month forecast

The graphs shows weekly sales and the sales prediction in 5 months. • Sales started at a low point, around 2k in transaction amount. They experienced a significant increase, peaking at around 10k. After the peak, sales declined slightly but remained above the initial level. The blue line suggests the sales forecasting, as can be seen, the predicting sales is not very good, though it does increase, there gonna be a lot of dips if continue with this operation, we need to boost it significantly.

3. Product total sale

Read slides then: The restaurant’s sales are driven significantly by Sandwiches and Frankie, highlighting their popularity. Items like Panipuri and Vadapav might require marketing efforts or menu adjustments to boost their performance. The combination of snacks (Sandwiches, Frankie) and beverages (Cold Coffee, Sugarcane Juice) contributes significantly to overall revenue.

4. Items and time sales correlation

Sandwiches sales are the best across almost all time slots, with the largest contribution at night ($17,520). Besides, Frankie and Cold Coffee show consistent sales across time slots. The same thing happened with Vadapav and Aalopuri, which are all least favored in all day. As seen from the graph, the grey color, which is night sales, takes quite a lot of space, so we would take a look of the distribution of each time of the day sales

1. Time Sales distribution

Sales at night had the highest sales, with taking up 22.5% for the total average sales in a day, followed by sales in the afternoon with around 20.5%. Meanwhile, midnight had the lowest sales. This implies at night and afternoon, the store got more sales than in other time of the day. This could be because of people usually spend more money eating out at night and afternoon, the time they just finish their work for a break or after their shift so they want to fill their stomach full. During the midnight, there maybe less sales because people typically do not eat out during midnight hours as they are sleeping.

1. Transaction type distribution

There were more transactions paid by cash than online. However, the gap was not large since 54% transaction was paid by cash, and the other around 46 % of the total transactions was paid online. However, we can take advantage of this distribution to implement sales-boosting solution

**V. Recommendations**

1. Leverage Popular Items

- Promotions on Sandwiches and Frankie: Offer discounts, combo deals, or loyalty rewards for these top-performing items to encourage repeat purchases and attract new customers.

- Upsell Beverages: Pair top-selling snacks with beverages like Cold Coffee or Sugarcane Juice at a slight discount to increase average transaction values.

2. Enhance Underperforming Items

- Improve Marketing for Panipuri and Vadapav: Highlight these items in advertisements or introduce limited-time offers to boost interest.

- Menu Innovations: Introduce new variations or flavors of underperforming items to cater to broader tastes and trends.

3. Optimize Time-Based Sales

- Night Specials: Launch “Dinner Deals” or late-night promotions to even more boost the sales on the highest sales period

- Midnight Offers: Although customer in midnight is not as much as in other time, we can attract late workers with this offer, provide niche promotions like discounts for delivery or snacks tailored to late-night cravings.

4. Leverage payment preferences:

As shown in the latest graph, customers tend to use cash to pay their transactions. So we can offer small discounts or cashback for online payments to encourage digital transactions while maintaining cash-friendly options.

5. Focus on Predictive Analytics

- Sales Trend Adjustments: Use insights from forecasting data to prepare for dips and boost sales through timely campaigns or menu adjustments.

- Monitor Inventory and Demand: Optimize supply levels for popular items to avoid shortages and reduce waste for less popular ones.

**VI. Build machine learning models (shown in notebook)**

I want to build a pure time series models so I just use the transaction\_amount and date features in the model.

a) Train an ARIMA model



1. Train an Exponential Smoothing model



1. Train a Prophet model



1. Train a SARIMAX model



\* **RMSE Comparison**:

* ARIMA provides the lowest RMSE, indicating the best overall fit to the data.
* SARIMA follows closely with slightly higher error.
* Exponential Smoothing and Prophet have higher RMSE values, suggesting they do not capture the data as effectively.

\* **MAPE Comparison**:

* ARIMA also shows the lowest MAPE, reflecting a good balance between error and accuracy.
* SARIMA has a marginally higher MAPE than ARIMA but remains a strong candidate for seasonal data.
* Exponential Smoothing and Prophet have higher MAPE values, highlighting more significant discrepancies between forecasted and actual values.
* Clearly seen ARIMA has the best performance out of 4 models

A graph showing a graph of a graph

Description automatically generated with medium confidence

The graph compares actual values with forecasts from four models: ARIMA, Exponential Smoothing, Prophet, and SARIMA. Unexpectedly, it is not ARIMA but SARIMA appears to be the most accurate, closely following the actual values. Other models, while providing reasonable forecasts, show some deviation from the actual values.