

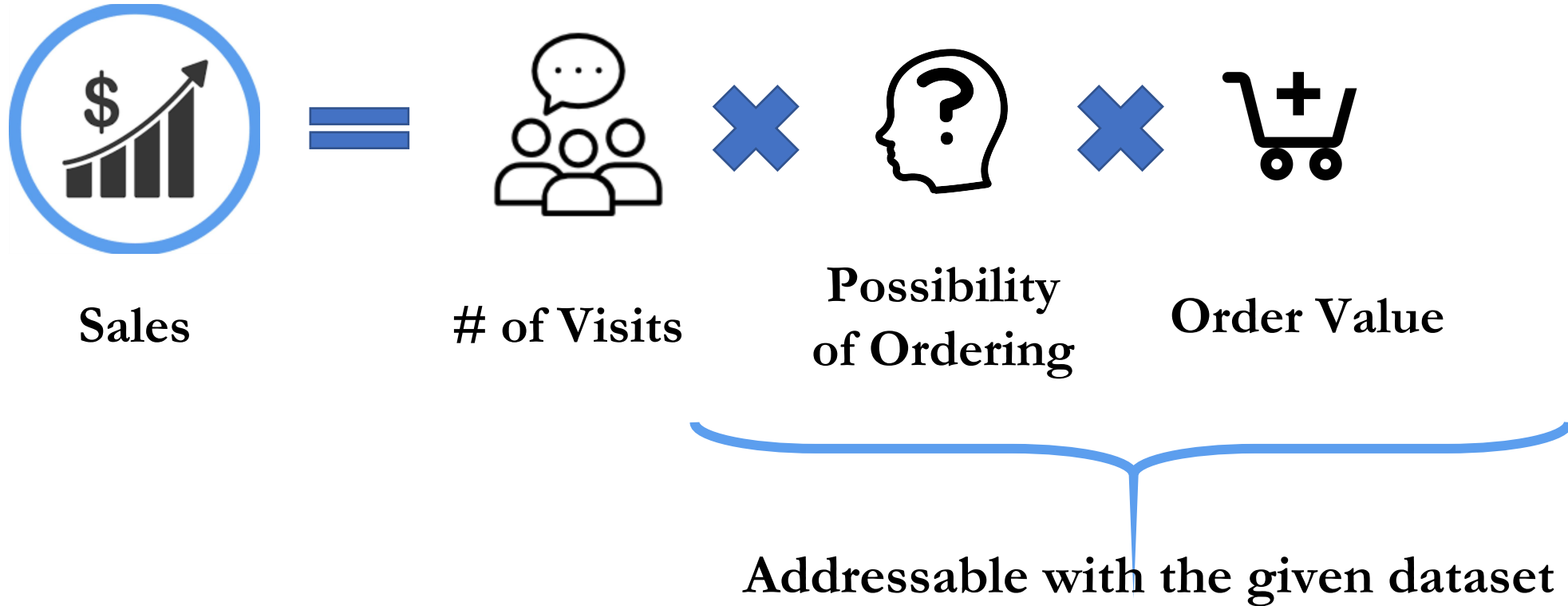


To Drive Up Overall Sales

Wayfair Datathon, Team 8:

Yu Tian, Wang Jin, Rui Wang, Hang Zuo

Logics to Drive Up Sales



Project Motivation

It is very common for e-commerce website meet the situation that people browse several products and random quit without any purchasing. Transferring the user browsing history to prediction of users' future clicking behavior will make huge impact on improving wayfair's business performance.

Two reasonable approaches to improve overall sales:

- **Finding the pattern of the user clicking behavior** will help with predicting either which kinds of page the user is heading to, or if the user is going to quit or not. Different strategies such as offering special discount or leading the users to more reasonable product sites can be applied to motivate the users' shopping behavior.
- **The second idea is to build general predictive regression model based on historical orders**, and find those important features which have great influence on the order value.



**Possibility
of Ordering**

Exploratory Data Analysis - Clickstream

Findings:

- The size of dataset of click_stream_with_purchase is 11057268 * 6 and the size of dataset of click_stream_without_purchase 11156464 * 5.
- Both dataset are cleaned and without missing values.
- It turned out there is significant difference in the average of length of sequential clicking behavior between two dataset. Users with purchases viewed 94.6 pages for a certain browsing while users without purchases only viewed 17.5 pages.



**Possibility
of Ordering**

Data Processing - Clickstream

- Filtered out users whose sequential viewed pages is below than 200 for the purpose of capturing the trend of user clicking behavior because we first want to explore user with more clicking behavior on wayfair.
- For generating training and testing samples, we used rolling windows to extract subsequence of user clicking stream with size of 20 and random step length within range from 1 to 5.



**Possibility
of Ordering**

Problem Definition- Clickstream

Input: 20 clicks behavior information before target click.

Output: The category of target click out of 25 total categories.

Model selection:

After our analysis, we consider that the information of the order of clicking behavior is very important. Recurrent neural network is proven to be very powerful capturing sequence information and we decided to use RNN model with LSTM cell to help with the classification task.



Possibility
of Ordering

Model Training - Clickstream

- We used a 80%-10%-10% train-validation-test split.

Dataset Size(# of sequences)		
Training Set	Validation Set	Test Set
391950	48994	48994

- Hyper-parameter Search

Hyper-parameter Range			Best Performance Combination		
# of Layers	Learning Rate	Hidden Size	# of Layers	Learning Rate	Hidden Size
1~3	0.0001~0.1	8~64	1	0.001	32

- Stop-criteria:

validation loss (patience = 5)



Possibility
of Ordering

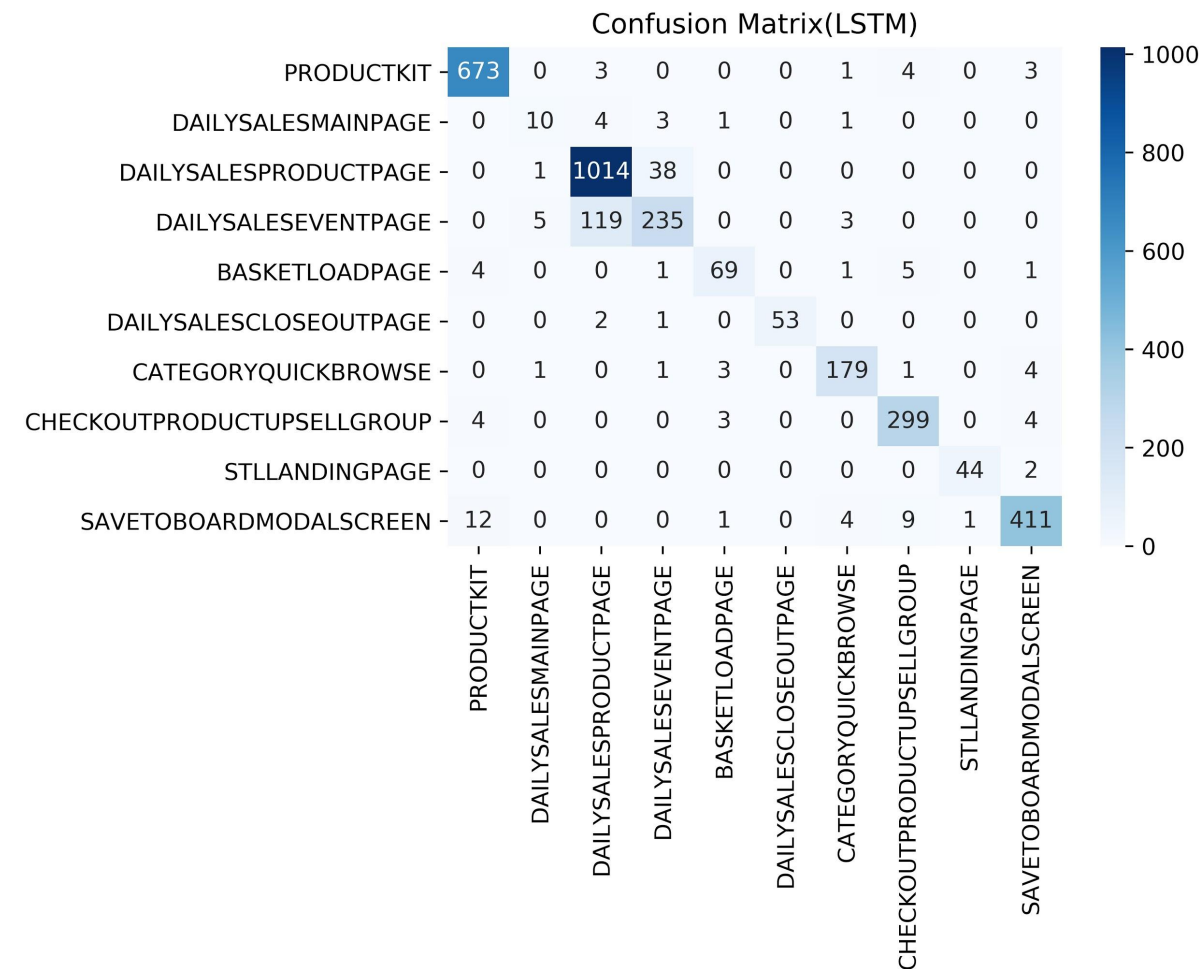
Model Performance - Clickstream

- Both metrics show promising results, which means behavior patterns of customers are predictable and can be captured by LSTM.

- Accuracy

Accuracy		
Training Set	Validation Set	Test Set
70.79%	71.40%	72.04%

- Confusion Matrix (Test Set)
(10 pages randomly selected)





Exploratory Data Analysis - Orders

Initial Thoughts:

- $\text{Order Value} = \text{Quantity of Order} \times \text{Onsite Price}$
- Possible to find the features that can significantly affect order value
- Intuitively related features:
 - Customer Income Level, Time of Ordering, On Promotion or Not
 - Item Price, Number of Reviews, Item Weight

Deliverables:

Business insight on what features to improve that could result to increased value per customer order.



Features Engineered for Regression Model

- Customer Income Level Estimation by Zip
 - Weighted Average of Adjusted Gross Income Size by Population
 - Join of Customer's Zip with Property Rental & Sales
- Day of the Week [0, 1, 2, .., 6]
 - Assumption: User behavior may vary across a week
- Time of the Day [1, 2, ... 24]
 - Assumption: User behavior may vary across a day
- Weight of the an item in its category
 - Assumption: For each category, customer may have different preference of weight, i.e., people like heavy sofa and light cook.



Features Engineered for Regression Model

- Length of Description [# of char]
 - Assumption: Detailed description may help and encourage people to buy
- Number of Reviews
 - Assumption: # reviews shows how the product's popularity
- Onsite Price
 - Assumption: Price will affect people purchase decision
- On Promotion [0 or 1]
 - Assumption: People may tend to buy when there is a promotion

Visualization - Date & Time vs. Order Value

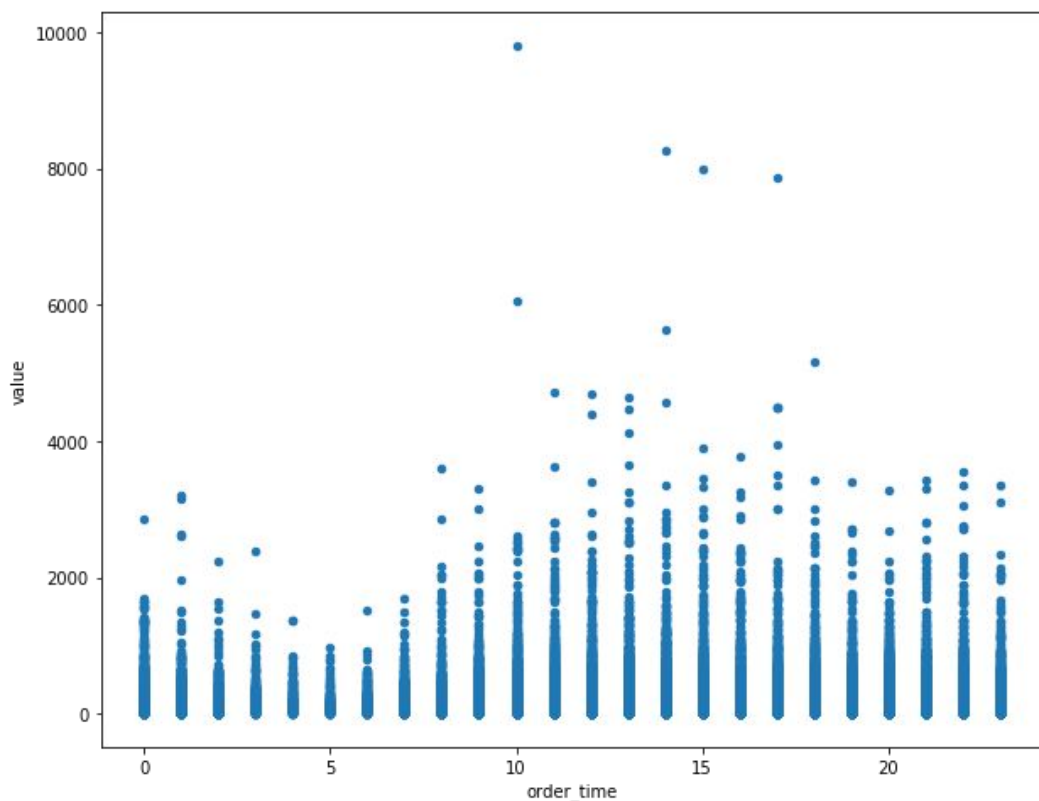


Order Value

(more EDA plots are in the Jupyter)

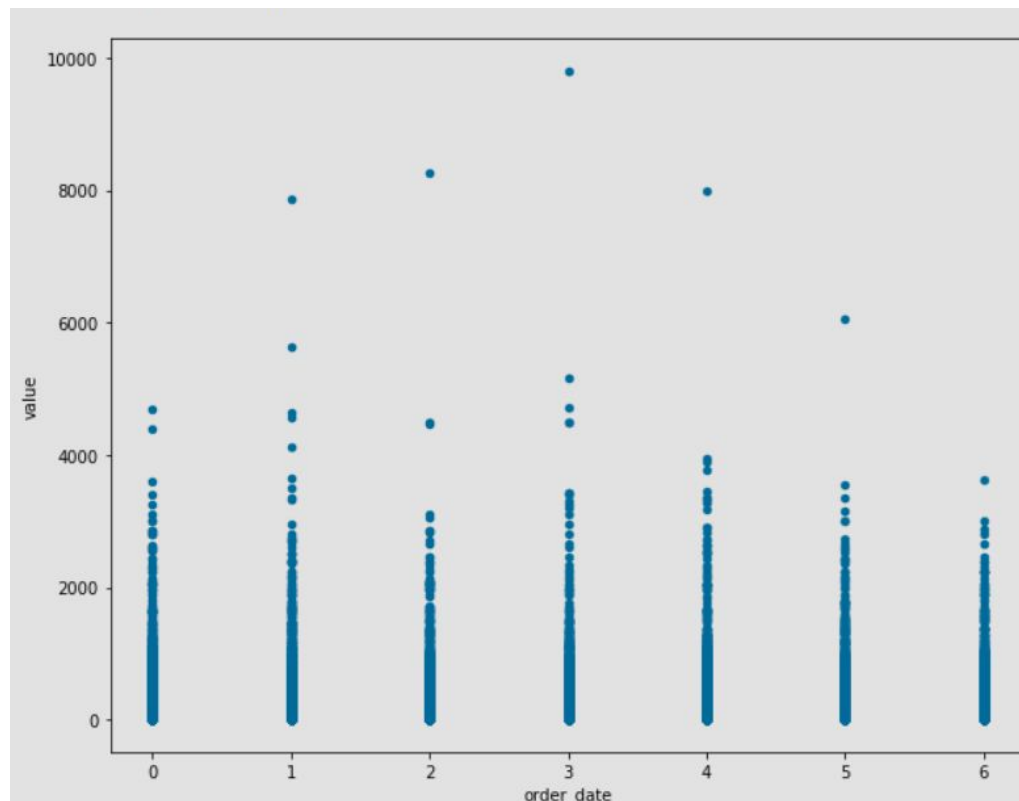
Ordering Time

Grouped by hours



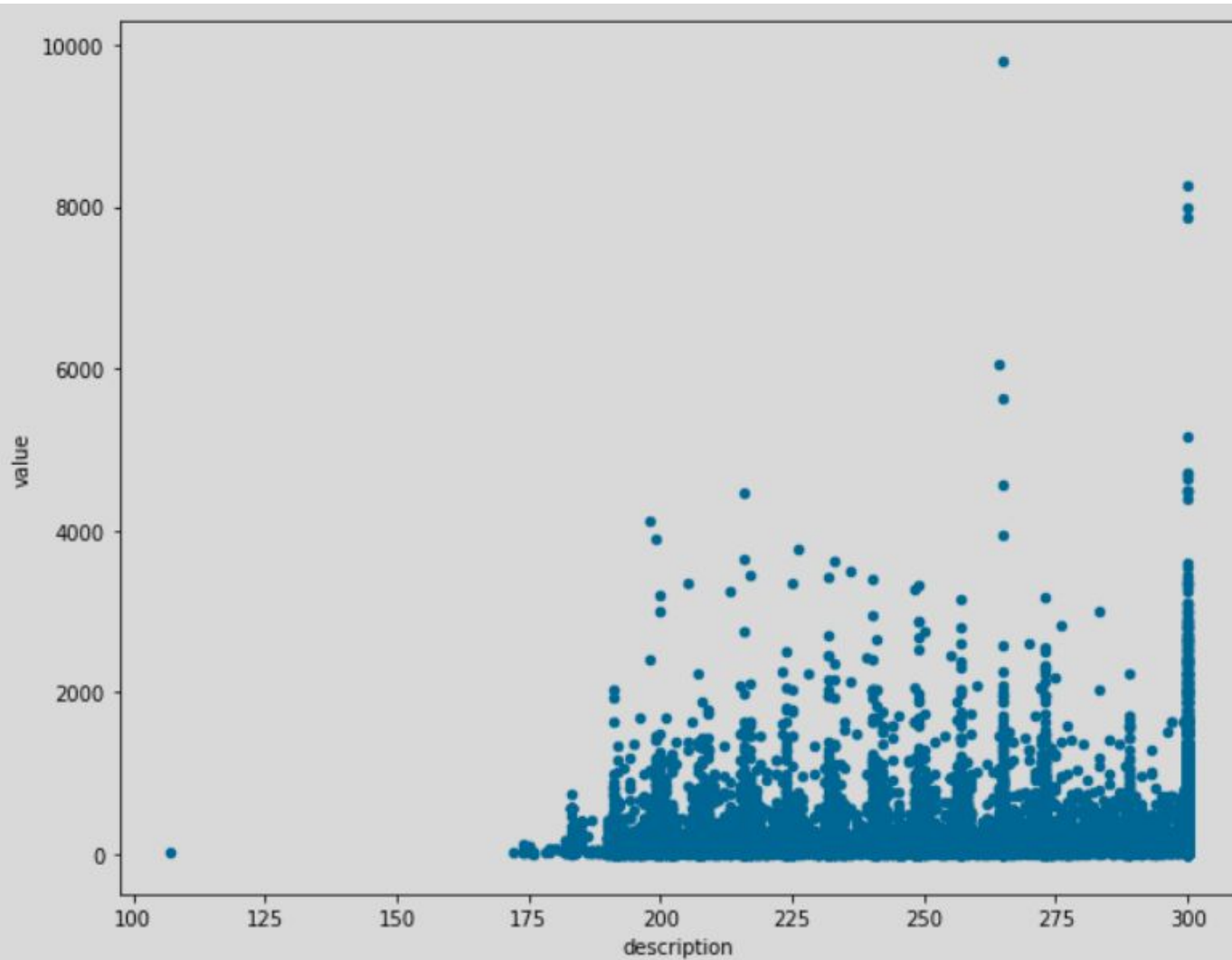
Ordering Day

of the Week





Visualization - Number of Reviews



Maximum number of reviews seem to be capped at 300, but the correlation still seems reasonable to further explore in regression model.

Significantly more higher outliers at higher # of reviews.

XGBoost Performance



Features	Importance
Order Date	0.077
Order Time	0.095
Length of Description	0.066
Item Weight	0.172
Onsite Price	0.289
# of Reviews	0.159
On Promotion or not	0.008
Customer Zip Income	0.134

Insights:

- Price & Weight are more linearly correlated to Value, thus used as benchmark for other features
- # of Reviews is highly important
- Customer Zip Income is important
- Current Promotions are highly ineffective in driving sales volume

Conclusion:

- It is proven possible to predict user's next click page category(with $> 70\%$ accuracy) as well as if the user will make the purchase or not during a certain browsing (based on if checkout page will appear in the sequence).
- Checkout / ordering page is within the prediction model
- Increase Number of Reviews could significantly increase sales
- Customizations by location could increase sales
- Current Promotions are significantly ineffective.

Feature Recommendation

Apply the predictive model of customer clickstream to detect when user increasingly demonstrate likelihood of Non-Buying, and provide customized Discount & Promotion before the predicted exiting.

Hypothesis for Testing:

Such feature could significantly increase sales and promotion effectiveness.