Time Series Sales Prediction For Walmart

Tiffany Chen, Amy Chiu,
Mandy Liu, Hyemin Yu

Agenda

01

Business Problem

Accurate Prediction for Daily Sales in next 28 days

02

Solution Map

Executive Summary

03

Exploratory Data Analysis

Finding Trends

04

Data Preparation

Preprocess Data for Model Building

05

Model Summary

LightGBM+LSTM

06

Business Value

Inventory Management, Resource Allocation

Business Problems (SCKQ)

Situation:

Predict most accurate daily sales of the next 28 days for Walmart.

Complication:

The Large Scale of Processing Data

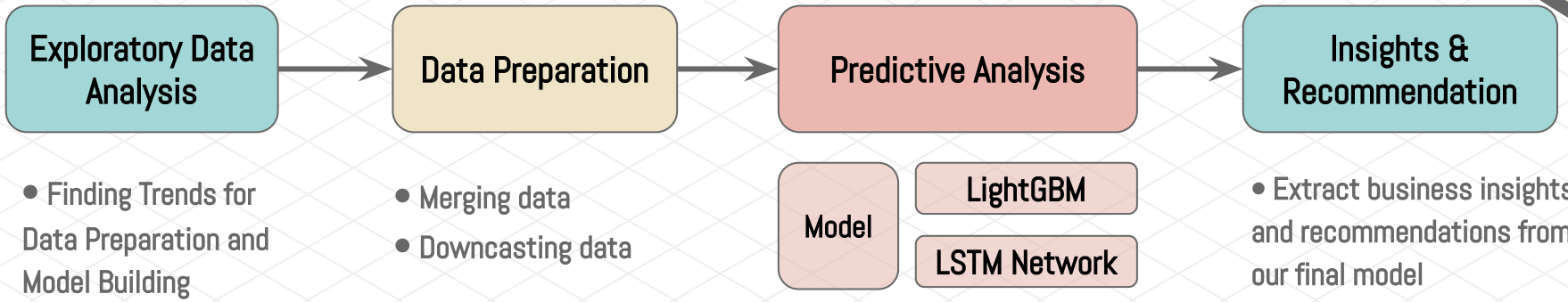
Feature Engineering and Feature Importance for Prediction

Different Prediction Strategy By State, Category, Department, Store

Key Question:

How can we develop forecasting methods that not only accurately predict 28-day ahead point and probabilistic forecasts for a extensive and hierarchical set of time series, but also ensure these methods are reproducible and scalable.

Solution Map



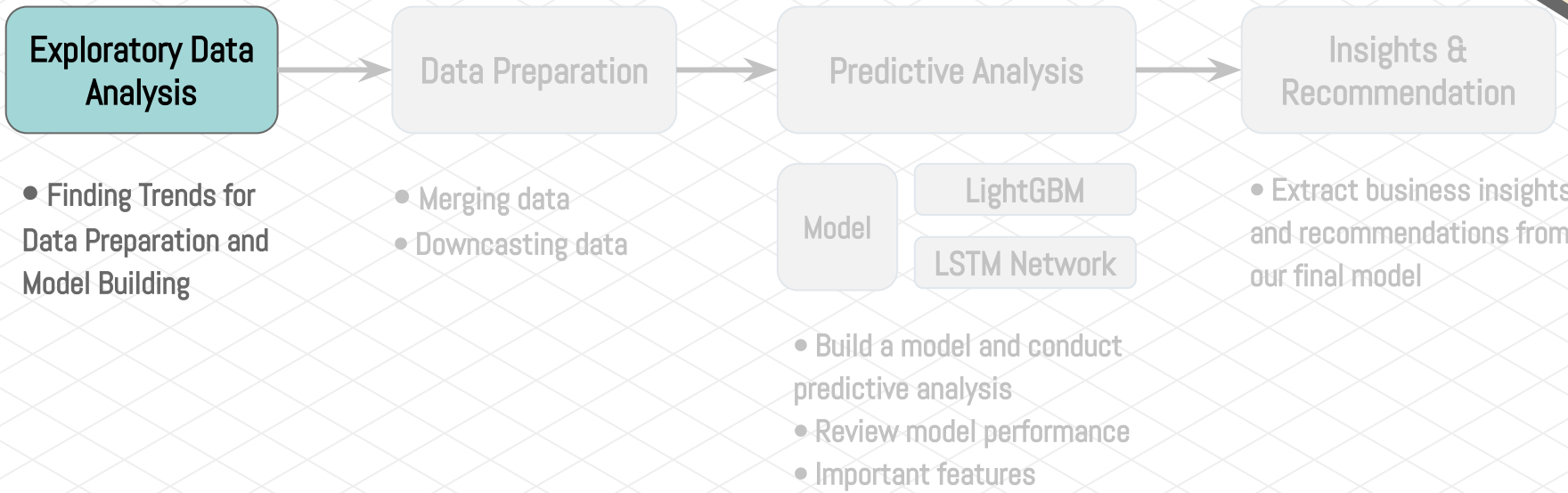
- Finding Trends for Data Preparation and Model Building

- Merging data
- Downcasting data

- Build models and conduct predictive analysis
- Review model performance
- Important features

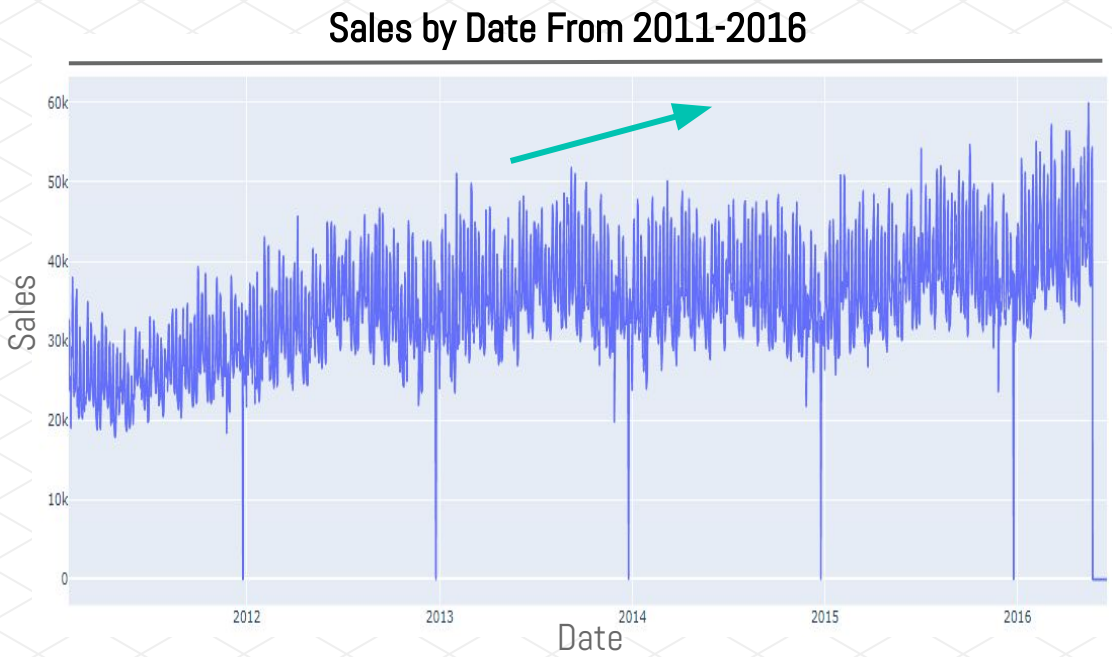
- Extract business insights and recommendations from our final model

Solution Map



EDA - Upward Trends in Overall Sales, Emphasized by Seasonal Patterns

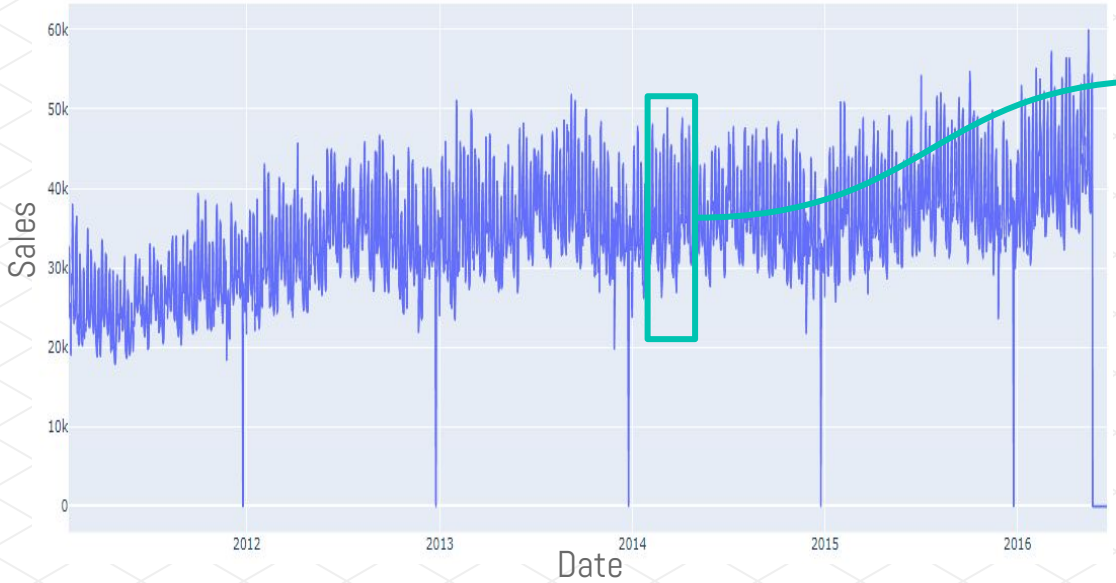
Models should incorporate sales seasonality



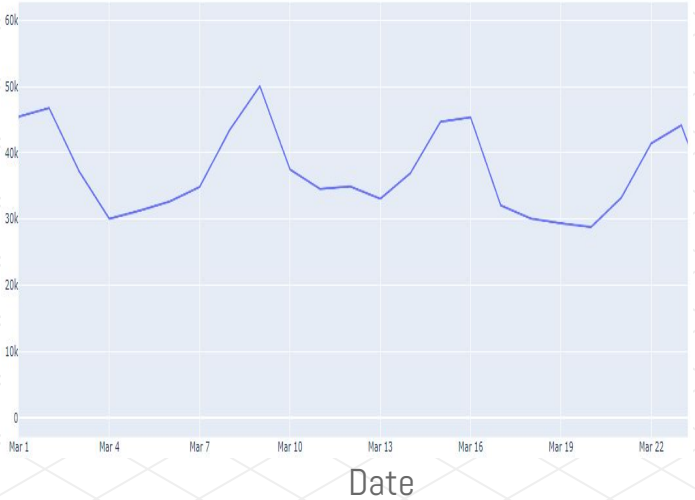
EDA - Upward Trends in Overall Sales, Emphasized by Seasonal Patterns

Models should incorporate sales seasonality

Sales by Date From 2011-2016

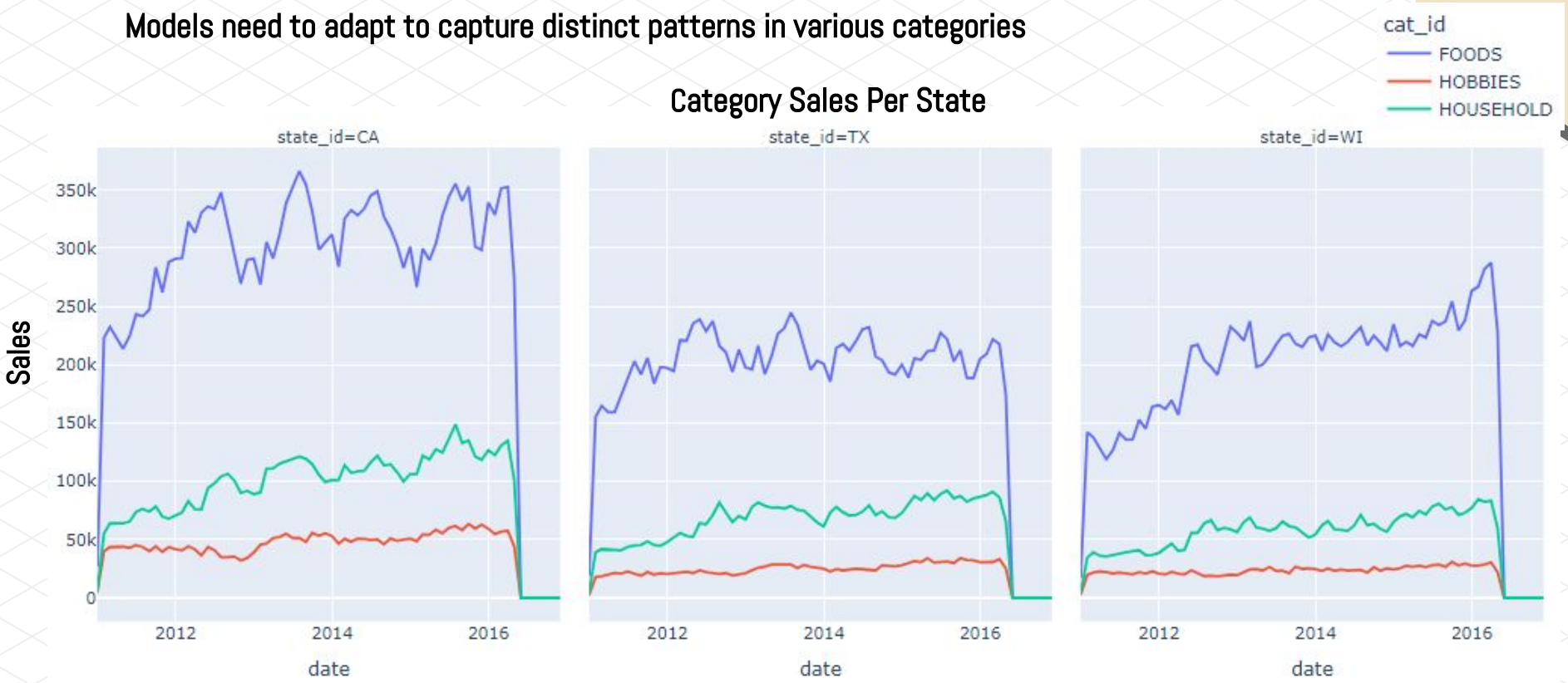


Weekly Seasonality Patterns



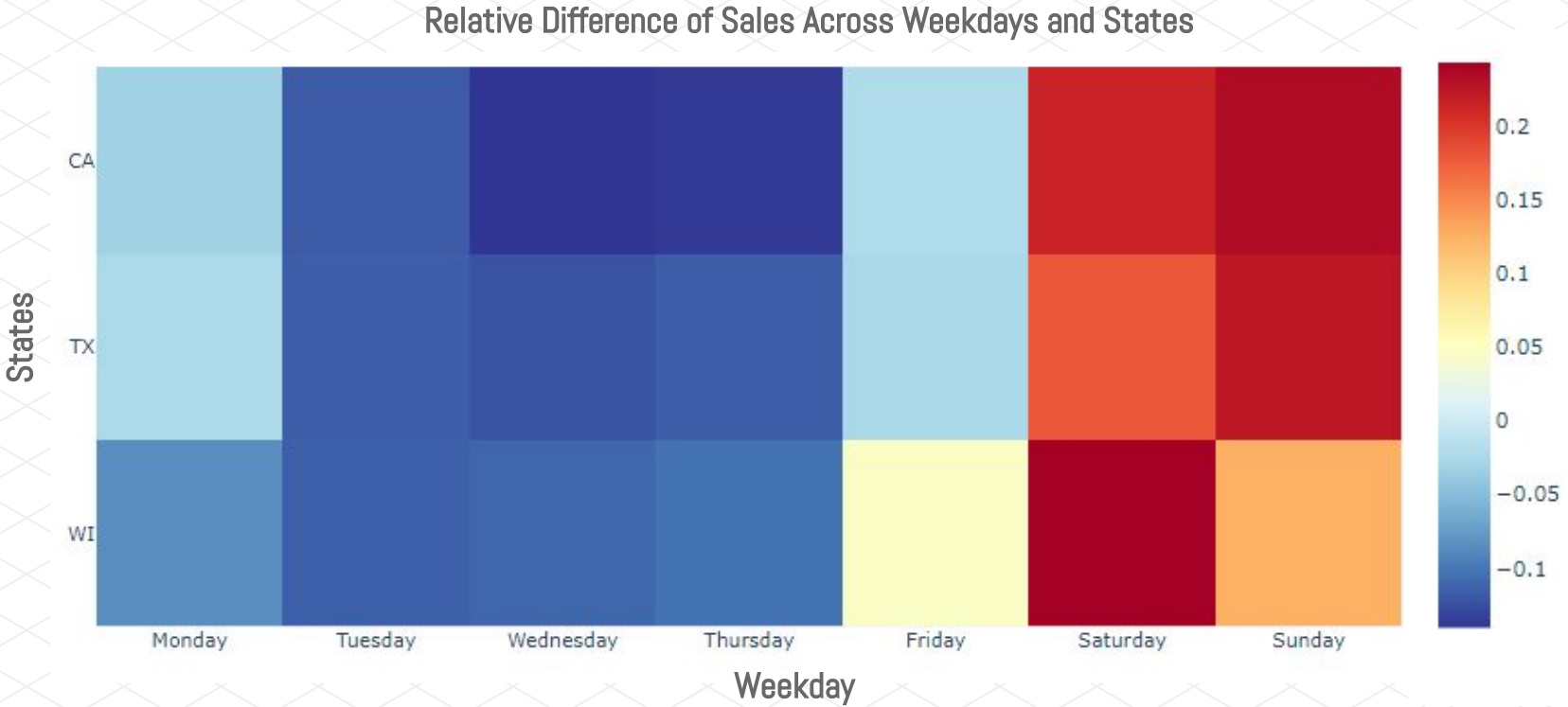
EDA - Sales Trends Vary Across States Over Time in Different Categories

Models need to adapt to capture distinct patterns in various categories

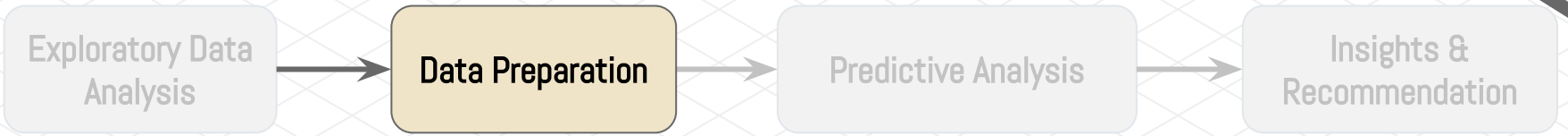


EDA - Noticeable Variations in Weekday Sales, Particularly Higher on Weekends

Relative Difference = (Weekday Sales - Week Average) / Week Average



Solution Map



- Finding Trends for Data Preparation and Model Building

- Merging data
- Downcasting data

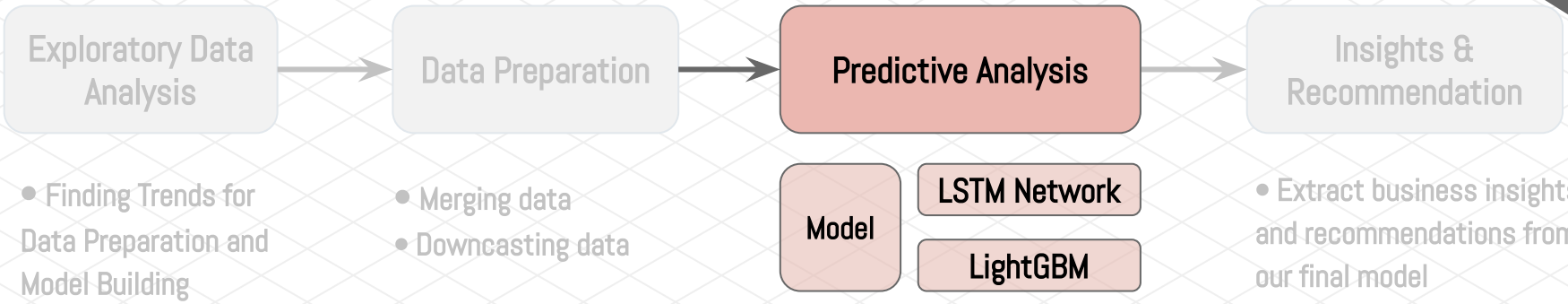
Model

- LightGBM
- LSTM Network

- Build models and conduct predictive analysis
- Review model performance
- Important features

- Extract business insights and recommendations from our final model

Solution Map



Model Summary

Goal

Multivariate Time Series Handling

Deal with complicated relationships

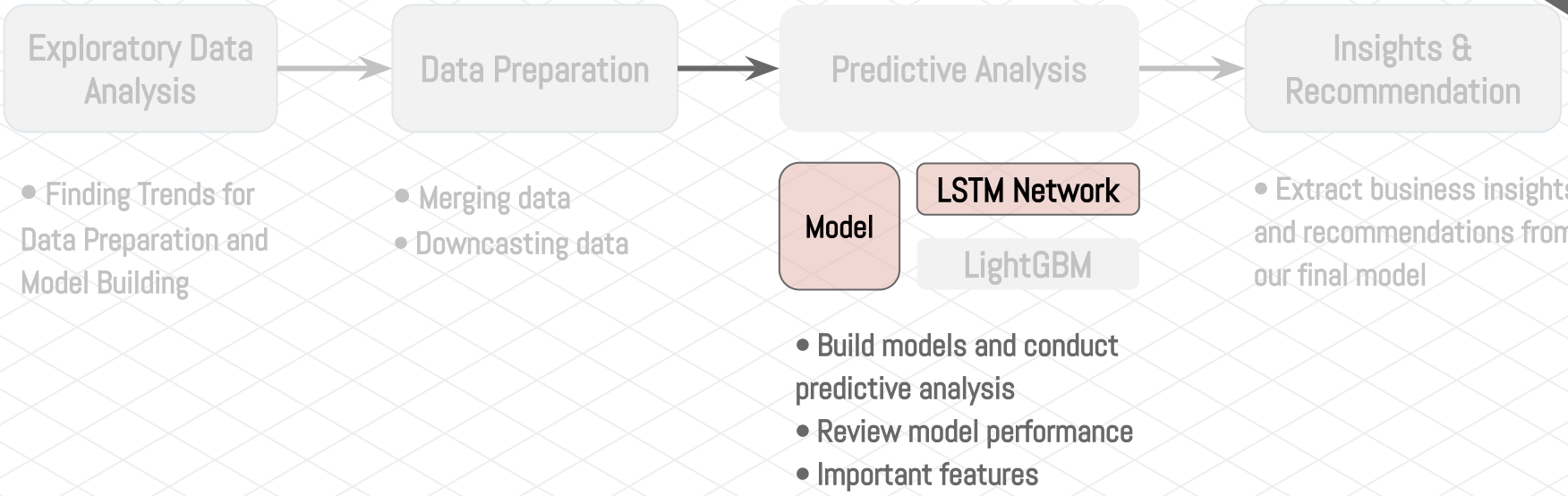
Ensemble Learning

Model Selection

Model 1: LSTM Network

Model 2: LightGBM

Solution Map



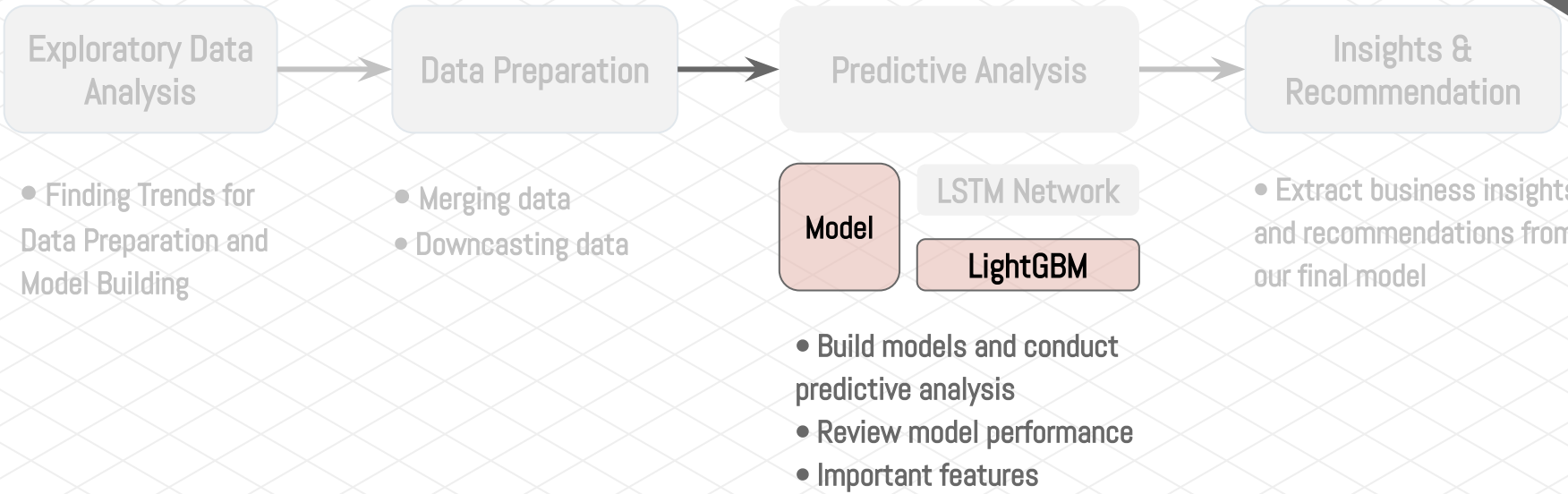
Model Summary - LSTM Network

LSTM: Capture long-term dependencies on sequential data where various factors such as Weekend and Event affect prediction

Structure	1. Define time range and start day based on trends analysis 2. Create feature to include impact of event and normalization 3. Hyper parameter tuning within various hidden layers 4. Use past 7 days (defined time range) to forecast the sales of the first unknown upcoming day
Scoring	Private: 0.6919 ; Public: 0.5461

D1908	D1909	D1910	D1911	D1912	D1913	D1914	D1915	D1916	D1917
Use these 7 days for prediction							Predicted value		
	Use these 7 days for prediction							Predicted value	
		Use these 7 days for prediction							Predicted value

Solution Map



Model Summary - LightGBM

LightGBM: Tree-based model for capturing complex interactions in sales data. It supports ensemble learning for enhancing the robustness and accuracy

Structure	1. Create features for capturing seasonal patterns based on trends analysis 2. Train separate models for different sets of store, category, department 3. Predict sales in 2-day intervals 4. Ensemble multiple models prediction results
Scoring	Private: 0.53199 ; Public: 2.56272

Why did we use Feature Engineering?

In Order To...	Used Features	
Capture Seasonal Patterns	Time Based	• day of week, month, quarter, or year • sales lag : 7/14/30 days
Smoothen Fluctuations	Rolling Statistics	• sales: rolling mean, exponential moving averages, rolling std. • price: rolling mean, rolling std.
Impact from Events	Time Since Last Event	• days elapsed since the last occurrence of SNAP event
Impact from Holidays	Label Encoding	• assigning a unique numerical value to each holiday
Stores / States / Category Differences	Mean Encoding	• average sales group by states / stores / categories / department / items

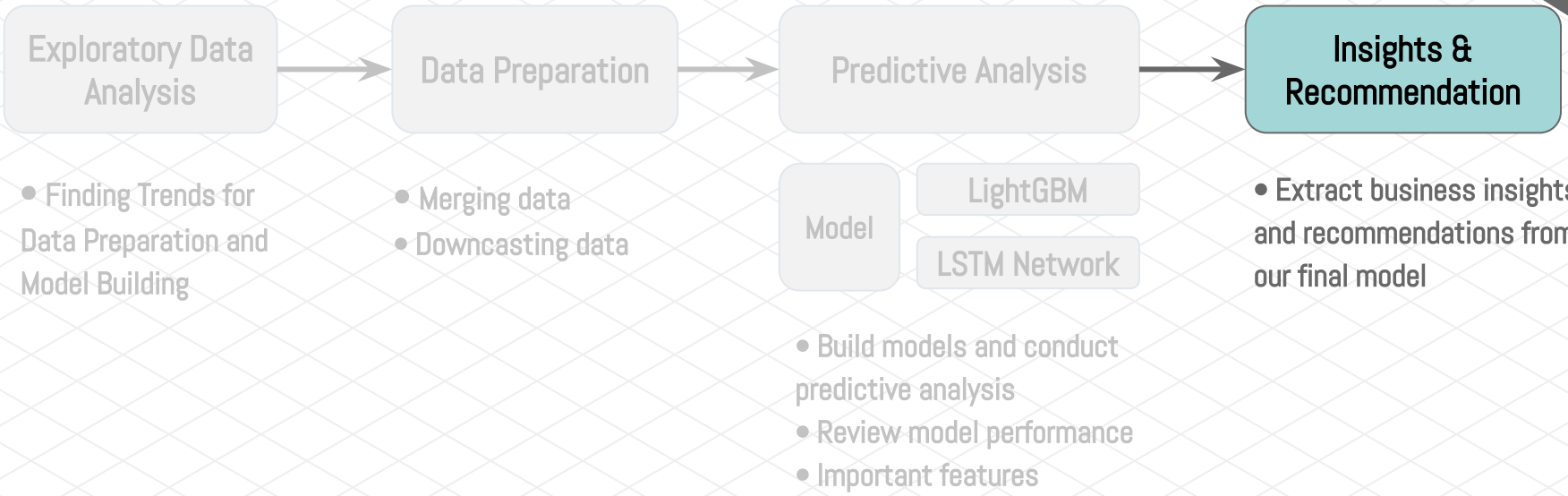
Model Summary

Goal	Model Selection	Scoring
Multivariate Time Series Handling	Model 1: LSTM Network	0.6919
Deal with complicated relationships	Model 2: LightGBM	0.53199
Ensemble Learning		

Model Summary

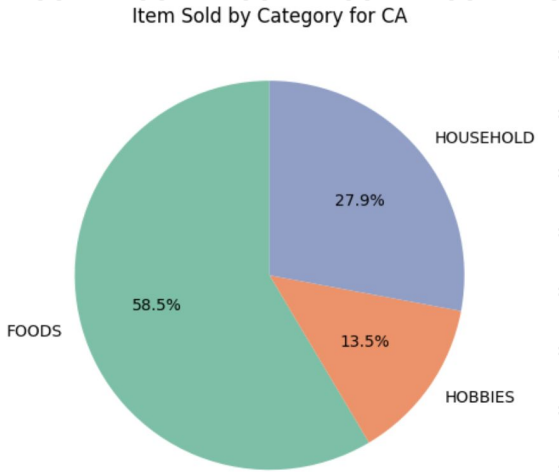
Goal	Model Selection	Scoring
Multivariate Time Series Handling	Model 1: LSTM Network	0.6919
Deal with complicated relationships	Model 2: LightGBM	0.53199
Ensemble Learning		

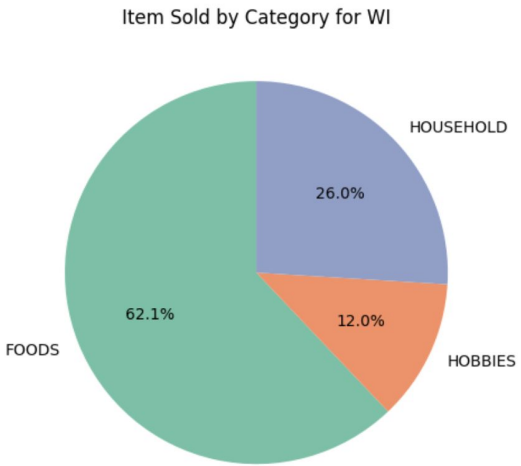
Solution Map

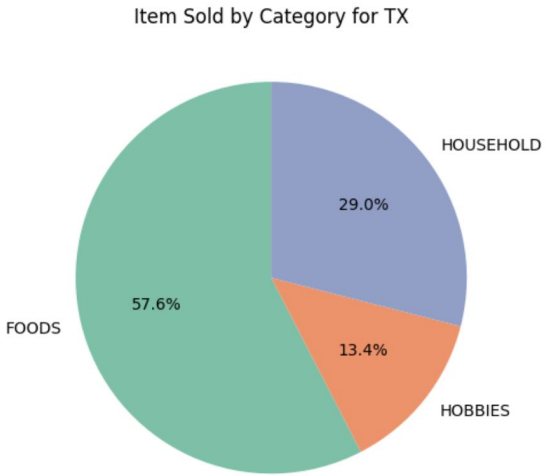


Insight & Recommendation

Based on our prediction, we can extract item sold distribution by categories in each state for the following 28 days







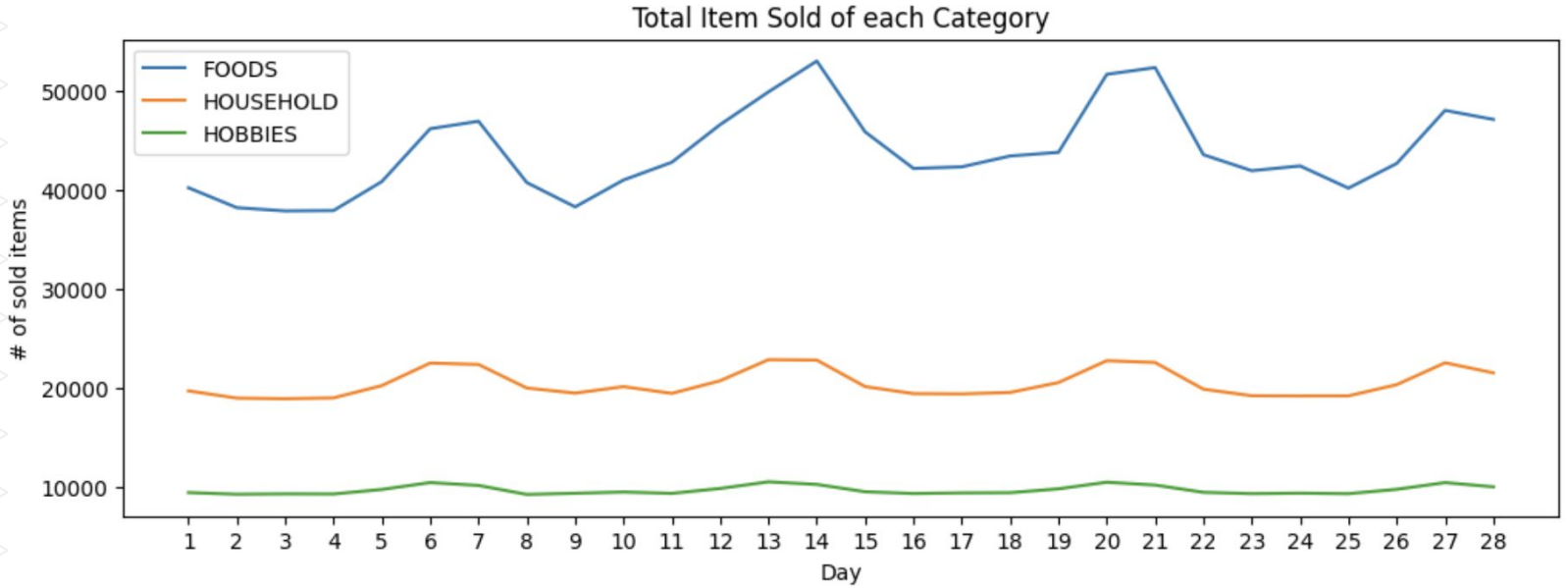
Insight & Recommendation

Identify number of sales by store in each state to optimize inventory allocation and streamline logistics management

State	Store (Order by # of sold)	Percentage
CA	CA_3	30.53%
	CA_2	25.45%
	CA_1	25.24%
	CA_4	18.78%
TX	TX_2	34.56%
	TX_3	34.19%
	TX_1	31.25%
WI	WI_2	37.47%
	WI_1	31.29%
	WI_3	31.24%

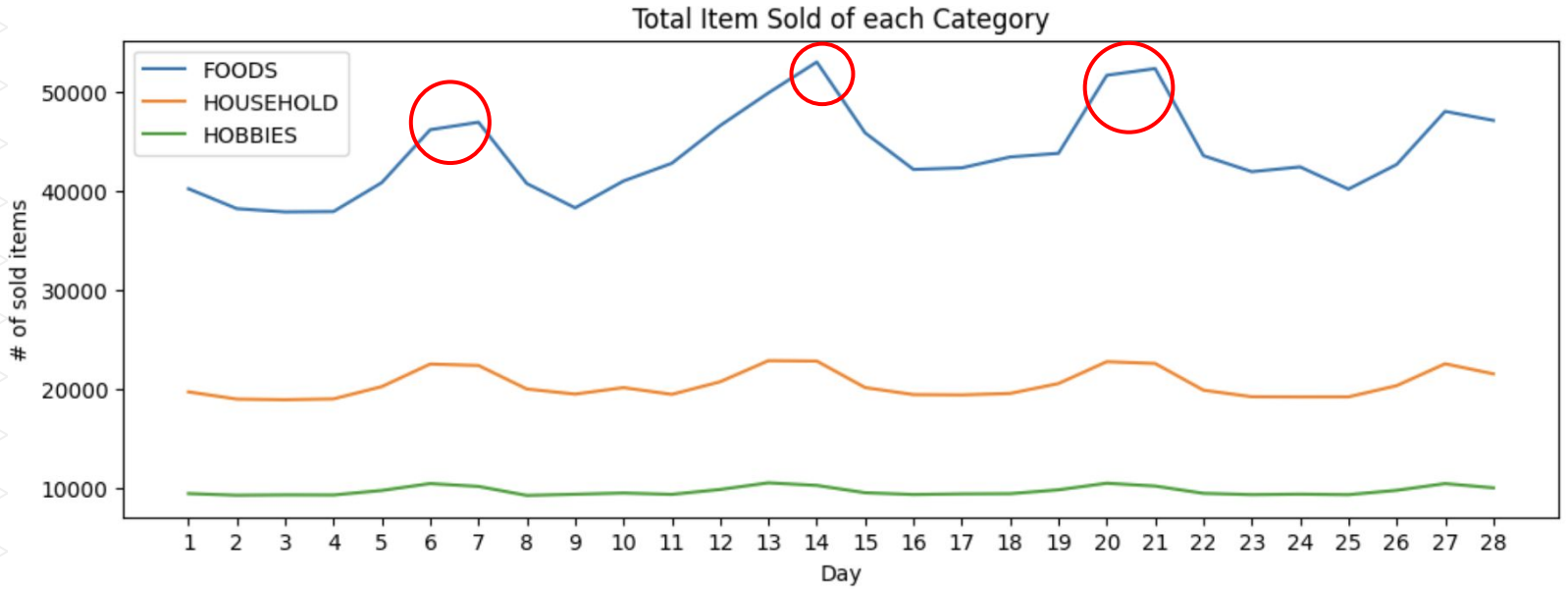
Insight & Recommendation

We can also analyze the patterns and peak of items sold by category over the next 28 days



Insight & Recommendation

We can also analyze the patterns and peak of items sold by category over the next 28 days



Thank You!