Sales Prediction For Walmart Retail Goods

MSBA 6421 - Team Project

Shivani Vats, Trung Le, Vineeth Gorla,
Rithvik Bhonagiri, Yufei Shen

December 5th, 2023

Prediction can bring business value to Walmart

Walmart's challenge in forecasting

- Inaccurate predictions can result in substantial losses for Walmart
- Traditional prediction methods are outdated, more new techniques are necessary
- Avoid costly mistakes and improve forecasting accuracy

Benefits of sales prediction

- Efficient Inventory Management
- Customer Satisfaction
- Smart Promotions
- Competitive Edge
- Optimized Supply Chain
- Support For Strategic Decision
- Reduce Financial Risks
- Raise Shareholder Confidence
- ……

Prediction can bring business value to Walmart

Situation

Walmart faces the task of maximizing decision-making efficiency with a rich dataset. Precise sales predictions becomes crucial to steer clear of both real and missed revenue opportunities

Key Question

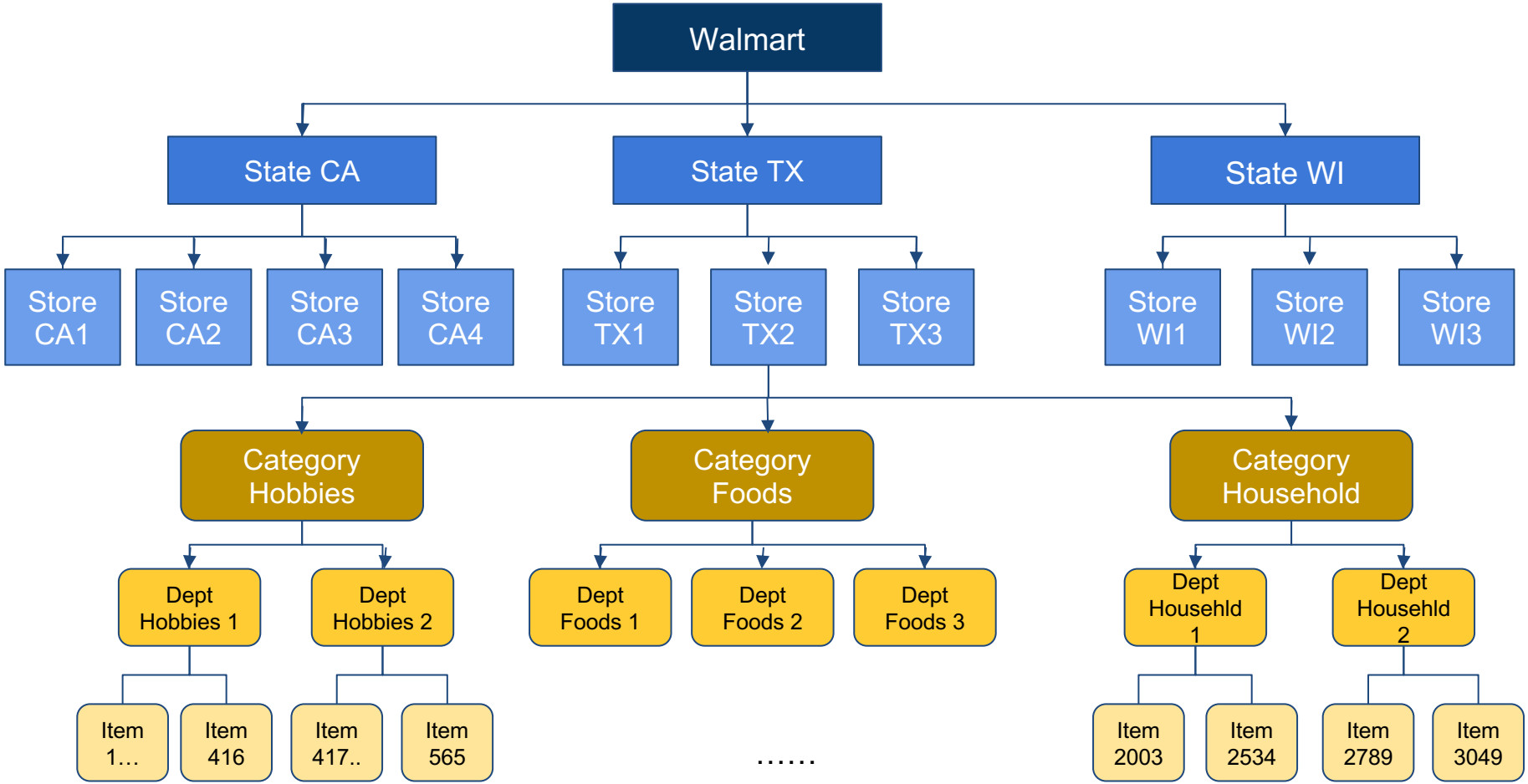
How to forecast daily sales for the next 28 days by using hierarchical sales data from Walmart?

Solution

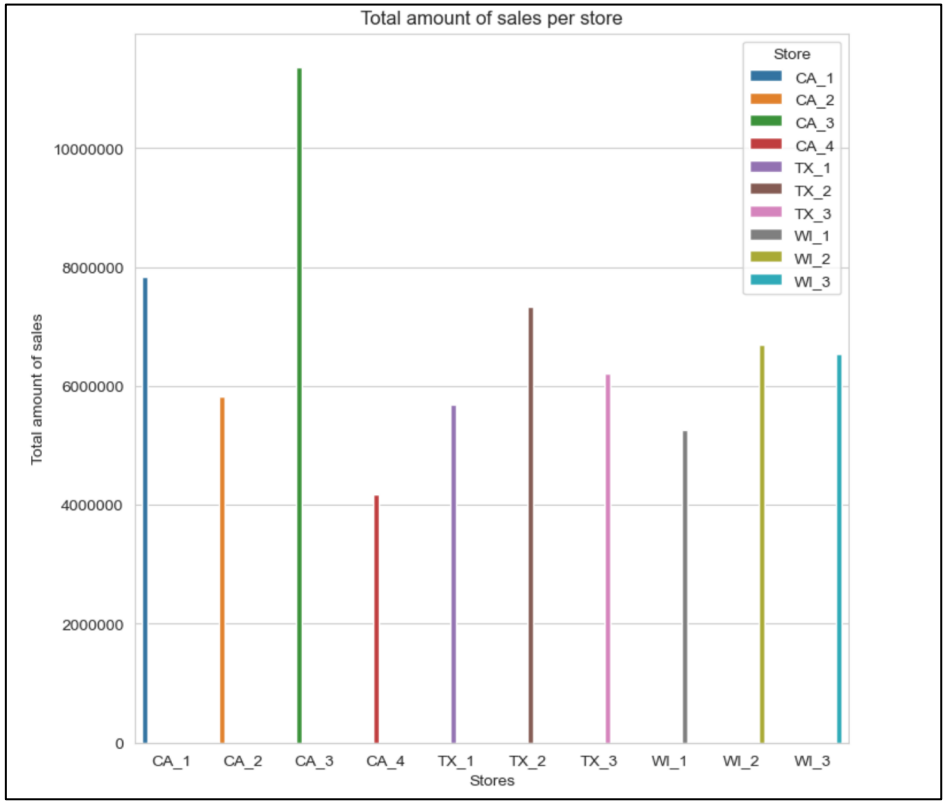
Leverage machine learning for predicting future sales, enhancing forecast accuracy in the process

Exploratory Data Analysis

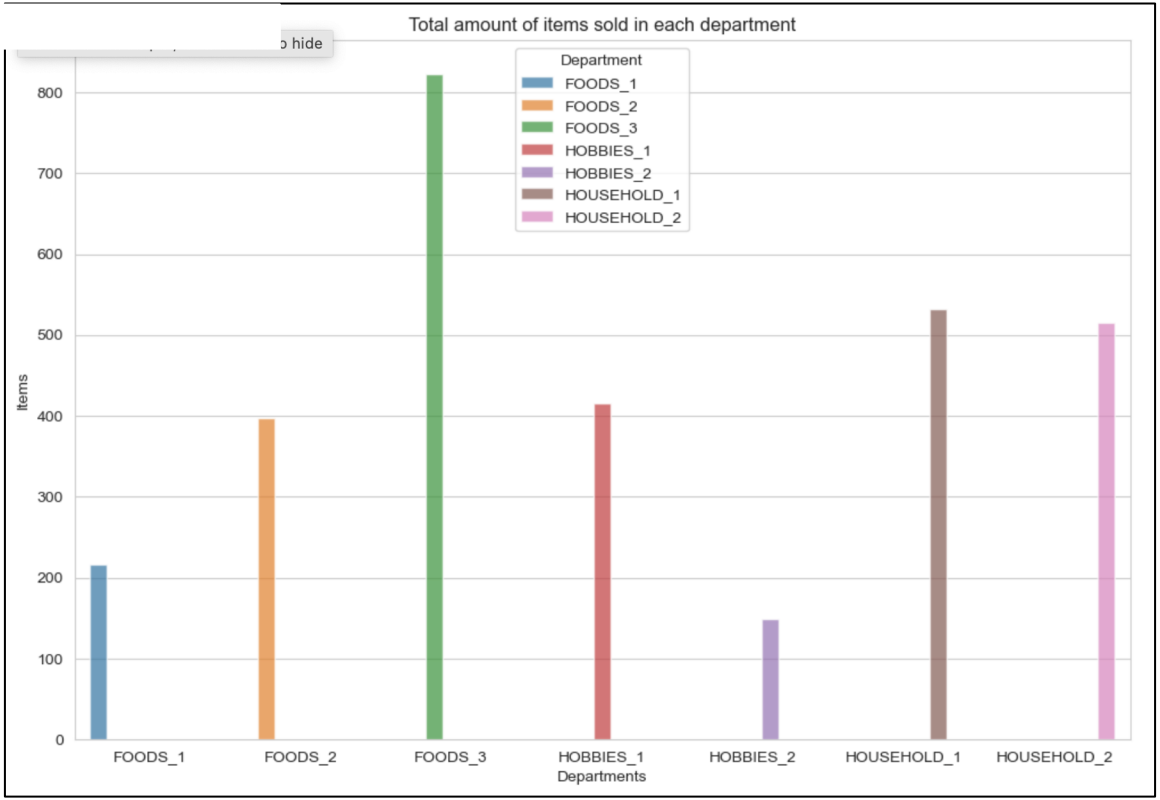
Raw Data	Description	# Feature	# Record	Data Size
calendar.csv	Workday & Special event day (e.g. SuperBowl)	14	1.9 K	103 kB
sell_prices.csv	Price of the products sold per store and date	4	6.84 M	203.4 MB
sales_train_validation.csv	historical daily unit sales data per product and store [d_1 - d_1913]	1019	30.5 K	120 MB
sales_train_evaluation.csv	sales [d_1 - d_1941]	1047	30.5 K	121.7 MB


Walmart
State CA
State TX
State WI
Store CA1
Store CA2
Store CA3
Store CA4
Store TX1
Store TX2
Store TX3
Store WI1
Store WI2
Store WI3
Category Hobbies
Category Foods
Category Household
Dept Hobbies 1
Dept Hobbies 2
Dept Foods 1
Dept Foods 2
Dept Foods 3
Dept Househld 1
Dept Househld 2
Item 1…
Item 416
Item 417..
Item 565
……
Item 2003
Item 2534
Item 2789
Item 3049

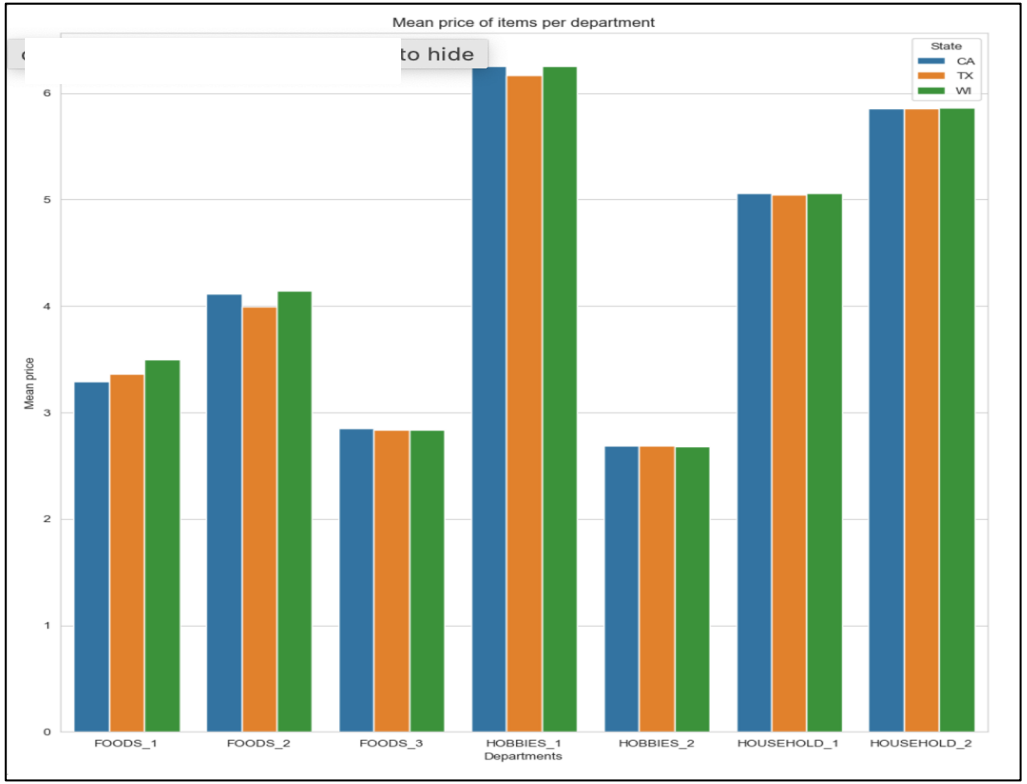
California has the maximum sales followed by Texas and Wisconsin



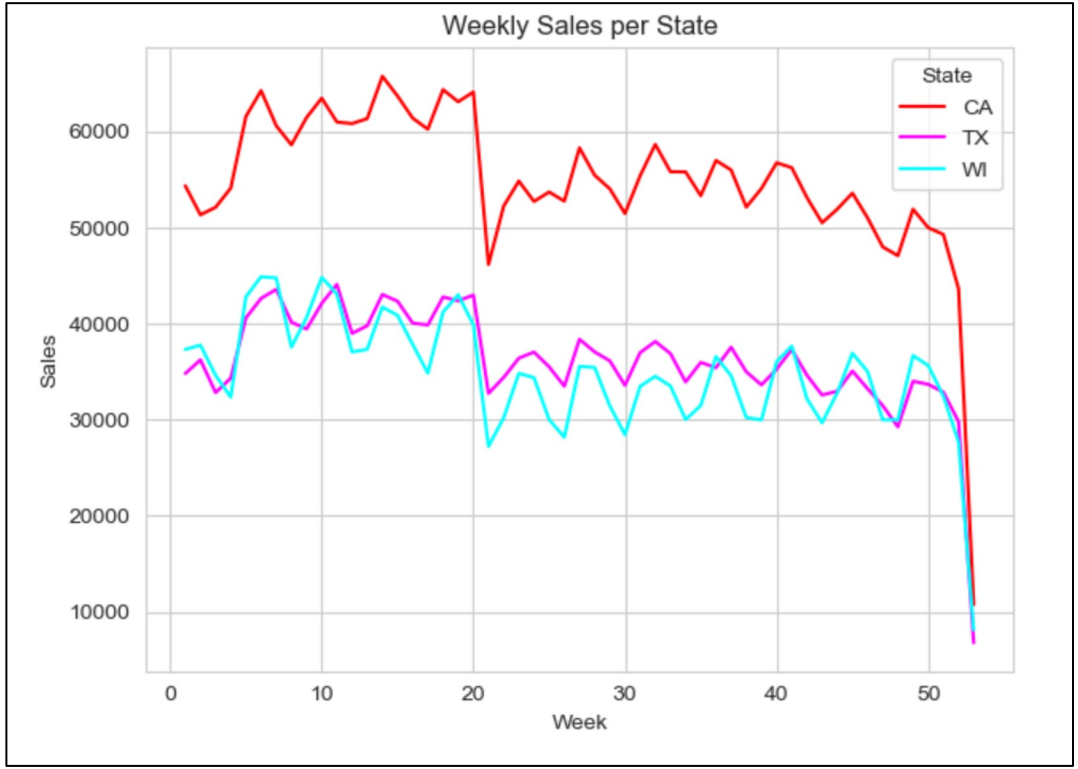
Food Category has the maximum sales



A sharp difference between Hobbies departments



Weekends are the most profitable days of volume sold



Feature Engineering

Calendar

Week days, months, years and events

Price

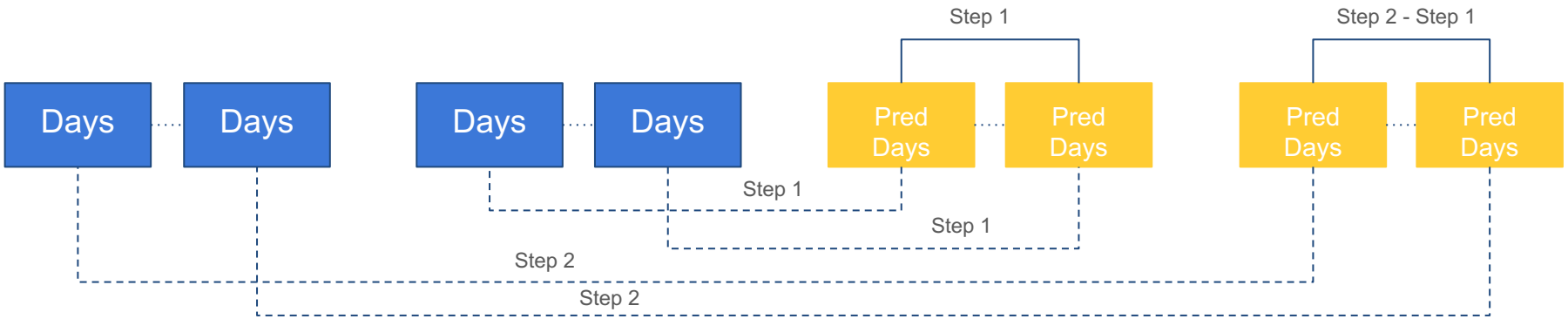
Expanding maximum, Expanding minimum, Expanding standard deviation of price, Expanding previous selling price

Sales

Leveraged Rolling Lags and Lags:
Rolling Lags -
a) Used window size *7, 14, 30, 60, 180 (In order to mimic weeks, months, year)*
b) Rolling zero *(it considers number of times sales were zero for a particular product)*
c) Rolling standard deviation
d) Rolling mean difference *(mean of sales difference per day in a particular window)*
e) Maximum sales in a respective rolling window

Lag Value of Sales:
a) Lag ranging between 0 to 15 *(to capture trends)*
b) Categorical encoding of items, department, category, stores
c) Expanding mean of sales group by category, item, dept, store, [store, category], [store, item id], and [store, dept id]

Modeling



<u>Example:</u>
If we consider stps of 14 and 28
Iteration 1:
Step 1 = 14
For 14 days predictions will refer back to T-14 days.
Iteration 2:
Step 2= 28
For next 14 predictions (which is 28-14), my prediction steps increases from 2
Model will refer back to T-28 days.

Leveraged models -

1. Light GBM
2. XGBoost
3. Neural Network Model
4. Ensemble model of these 3 models

Model Implementation

- Then we run prediction model where we first loop over store and department to train the model (slow) , next over store and category (will be quicker) and take average of both these methods to arrive at final submission.
- After features and methods were finalized, we decided to explore various regression algorithms.
- For the purpose of this project we have chosen , LightGBM, XgBoost, Neural network, stacked model of lightgbm, xgboost, Neural Network regression models.
- From the kaggle scores we found Light GBM was performing better than other models.
- Hence, we chose to modify the prediction step size from [14, 28] to aggressive range of [7, 14, 21, 28] (compromising the speed) and again ran Light GBM and got the best kept score.

Result Interpretation

Submission and Description	Private Score
MSBA.002.UnitedTeam_LightGBM.csv Complete (after deadline) · 3m ago	**0.54505**
MSBA.002.UnitedTeam_NeuralNetwork.csv Complete (after deadline) · 3m ago	**0.728**
MSBA.002.UnitedTeam_xgboost.csv Complete (after deadline) · 3m ago	**0.5598**

M5 Forecasting - Accuracy

Estimate the unit sales of Walmart retail goods

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

0/1

You selected 0 of 1 submissions to be evaluated for your final leaderboard score. Since you selected less than 1 submission, Kaggle auto-selected up to 1 submissions from among your public best-scoring unselected submissions for evaluation. The evaluated submission with the best Private Score is used for your final score.

Submissions evaluated for final score

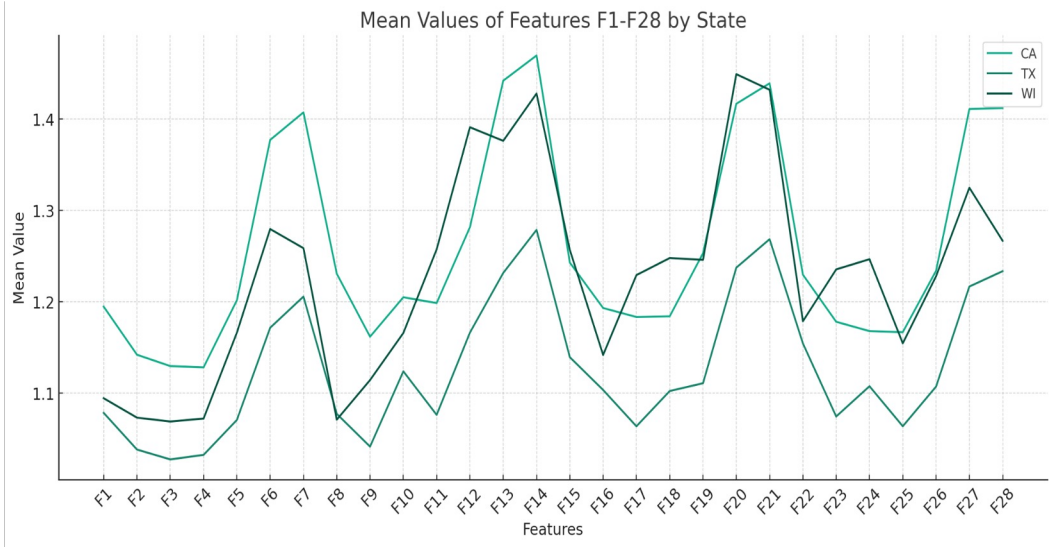
All Successful Selected Errors

Recent

Submission and Description	Private Score	Public Score	Selected
MSBA.002.UnitedTeam.LightGBM_best.csv Complete (after deadline) · 4m ago	0.53402	2.56272	☐
MSBA.002.stacked_predictions_NeuralNetworklightgbmxgb.csv Complete (after deadline) · 27m ago	0.58284	2.56272	☐

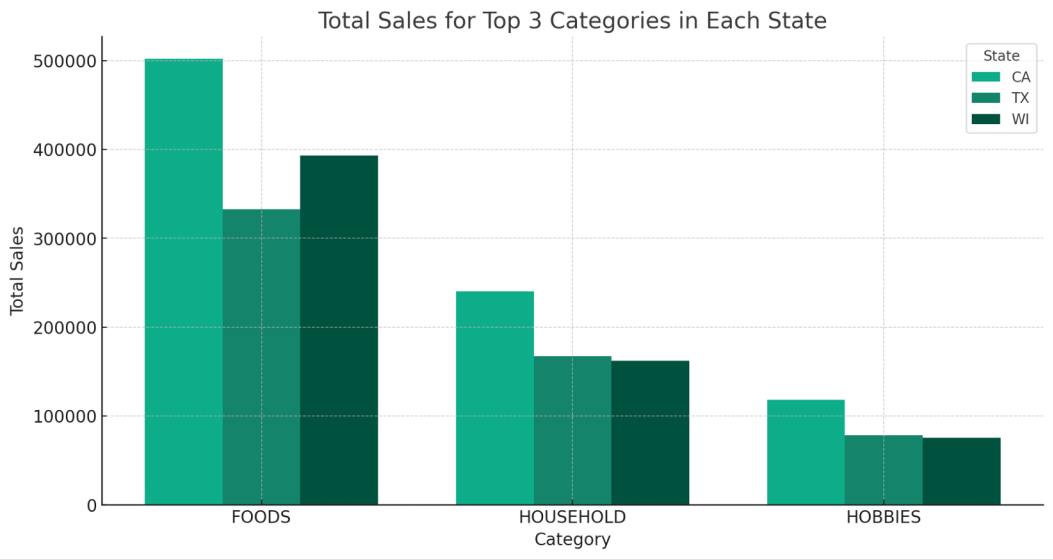
Findings from prediction result

- In terms of average sales
- Sales will be highest in California and lowest in Wisconsin



Findings from prediction result

- In terms of total sales
- Sales will be highest in Foods and lowest in Hobbies
- Wisconsin has high food sales, low hobby sales



Thank You!