A survey of ML Based sales prediction

Team: Hinton's Heuristics

Members:
JatinKumar Janshali (24M2021)
Prashik Patil (24M2014)
Shoaib Ahamed (24M2102)

Course Instructor:
Prof. Preethi Jyothi

Task Description

Aim:	• Predict Walmart's sales for the next **28** days using the **M5** dataset.
Dataset:	• **6 years** of Walmart sales data across **3 states.**
Method:	• Incorporate events, holidays, and promotions into the forecast.
Focus:	• Evaluate **Linear Regression**, **LSTM** and **LightGBM** models.
Evaluation:	• The RMSSE formula from the M5 competition is the following:

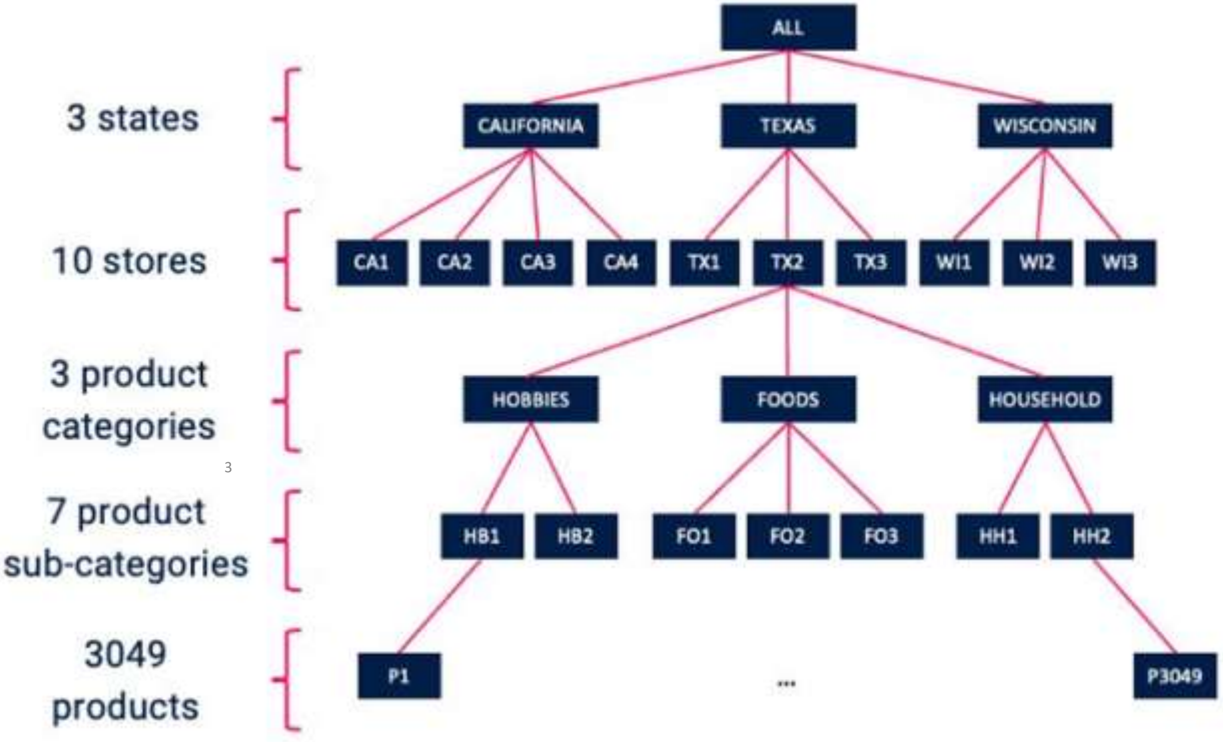
$$\boldsymbol{RMSSE} = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - \hat{Y}_t\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t - Y_{t-1})^2}},$$

Dataset Details

Calender
- Holiday, Promotion etc details
- 215 MB
- 1969x14

Sales_train
- Per day sales, item wise
- 452 MB
- 30490x18

Sell Prices
- Item wise sale prices
- 208 MB
- 6841121x4



Data Preprocessing

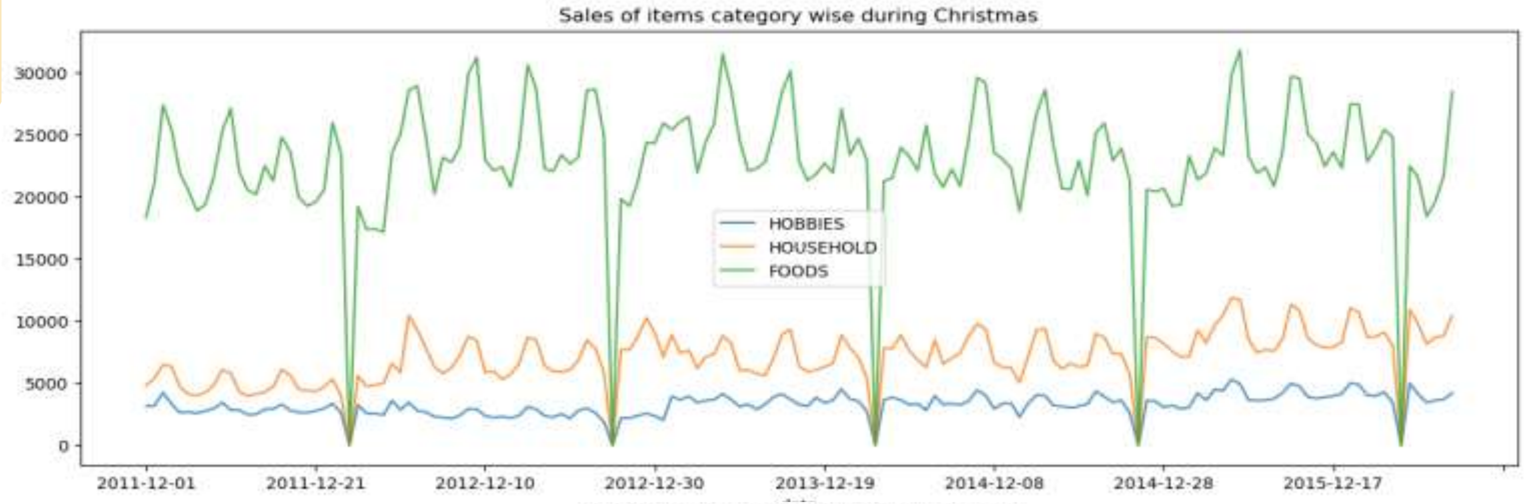
Memory Optimization with Down casting

- **Purpose:** large data types (**float64**) into smaller ones (**float32**) w/o losing precision.
- **Implementation:** Dynamically using NumPy.
- **Impact:** Optimized data storage and faster processing
- **Result:** Reduction in size upto 45% *.

Data Normalization with Min-Max Scaling

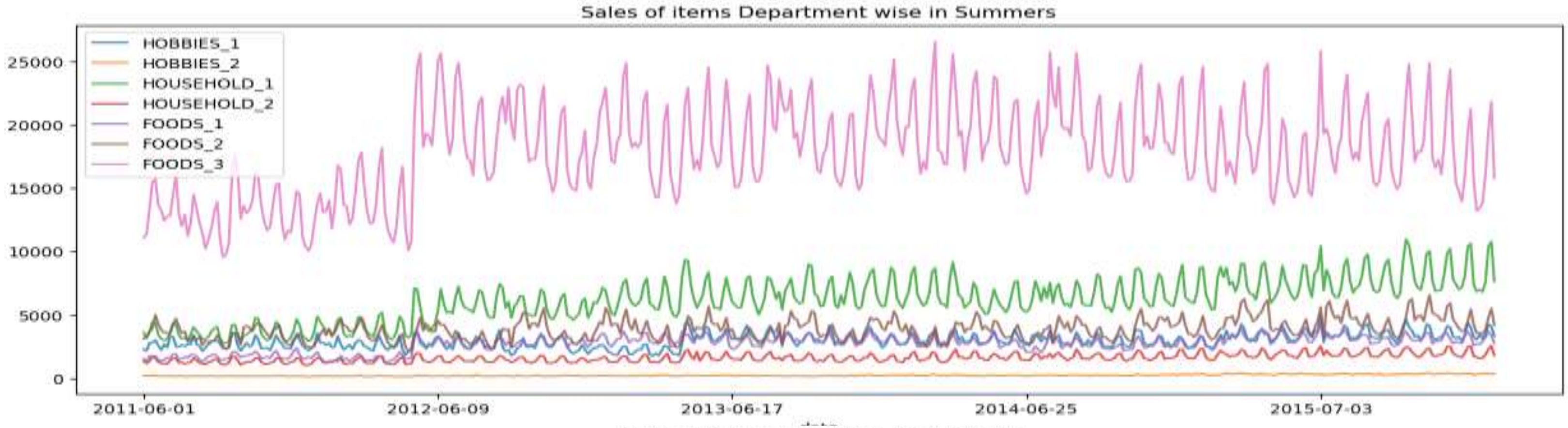
- **Purpose:** Standardize numeric data to a common scale (-1 to 1 in this case).
- **Implementation:** Utilizes **MinMaxScaler** while preserving the proportionality in features
- **Impact:** training stability and convergence for features with varying scales.

EDA

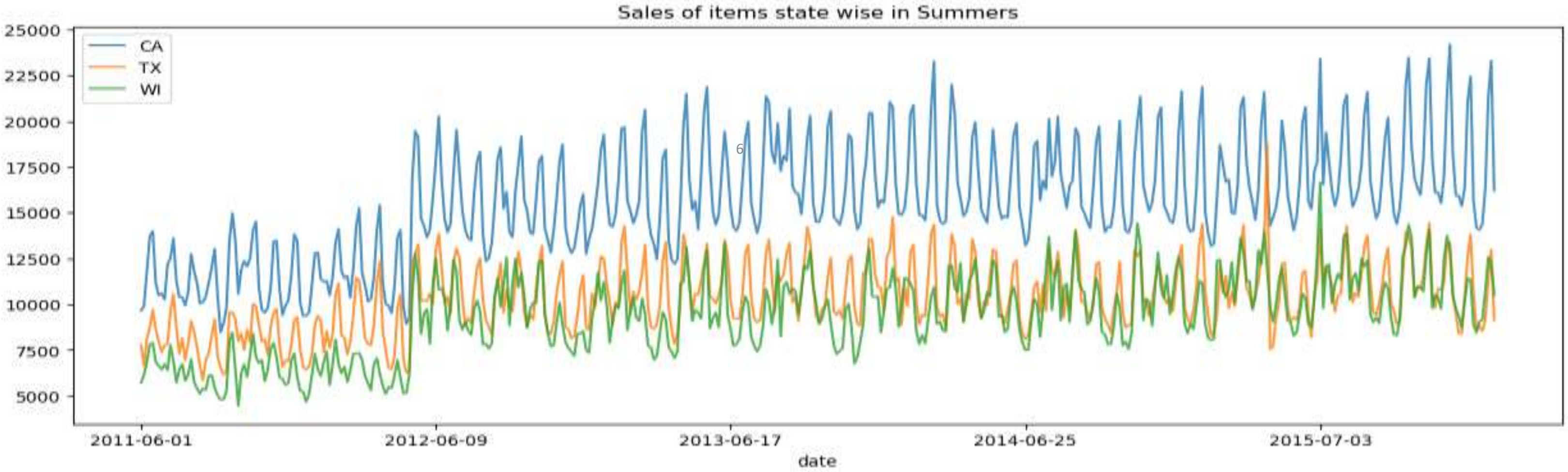

Sales of items category wise during Christmas
30000
25000
20000
15000
10000
5000
0
HOBBIES
HOUSEHOLD
FOODS
2011-12-01
2011-12-21
2012-12-10
2012-12-30
2013-12-19
2014-12-08
2014-12-28
2015-12-17


Sales of items state wise around Christmas
20000
15000
10000
5000
0
CA
TX
WI
2011-12-01
2011-12-21
2012-12-10
2012-12-30
2013-12-19
2014-12-08
2014-12-28
2015-12-17
date

EDA



Sales of items Department wise in Summers
HOBBIES_1
HOBBIES_2
HOUSEHOLD_1
HOUSEHOLD_2
FOODS_1
FOODS_2
FOODS_3
25000
20000
15000
10000
5000
0
2011-06-01
2012-06-09
2013-06-17
2014-06-25
2015-07-03



Sales of items state wise in Summers
CA
TX
WI
25000
22500
20000
17500
15000
12500
10000
7500
5000
2011-06-01
2012-06-09
2013-06-17
2014-06-25
2015-07-03
date

Linear Regression: Ridge regression model

Salient Features:

Tailored Predictions: Models can be optimized for specific category patterns and trends.

Reduced Noise: Less diverse data leads to better focus on relevant features.

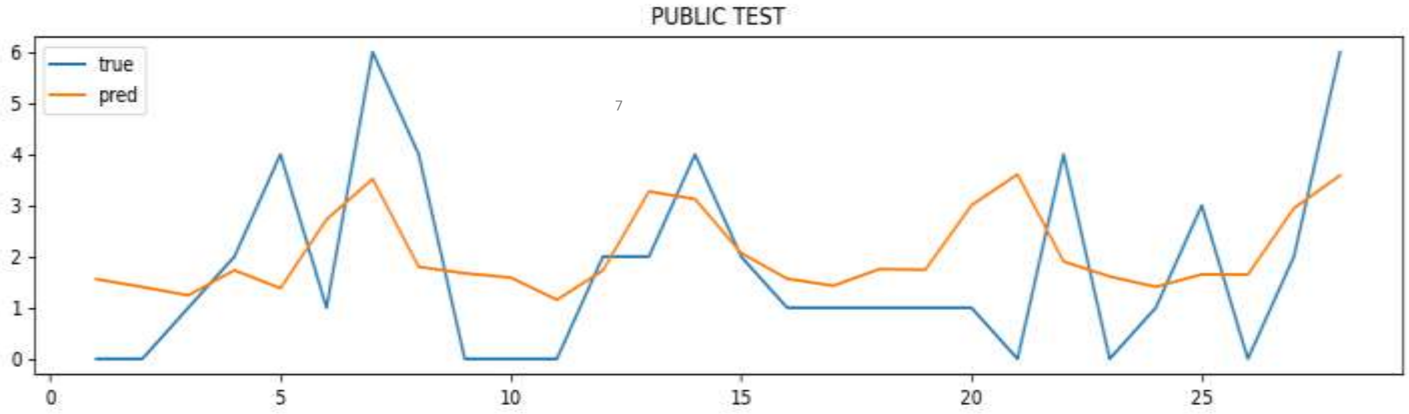
Faster Training and Inference: Smaller datasets accelerate model training and prediction.

Improved Interpretability: Insights from individual models can inform targeted strategies.

Enhanced Accuracy: Specialized models can yield more accurate predictions.

RMSSE Score - 0.69725

Sample Prediction for HOBBIES_1_004_CA_1



LSTM

Key Training Parameters

Number of Epochs: 500

Learning Rate: 1×10^−3

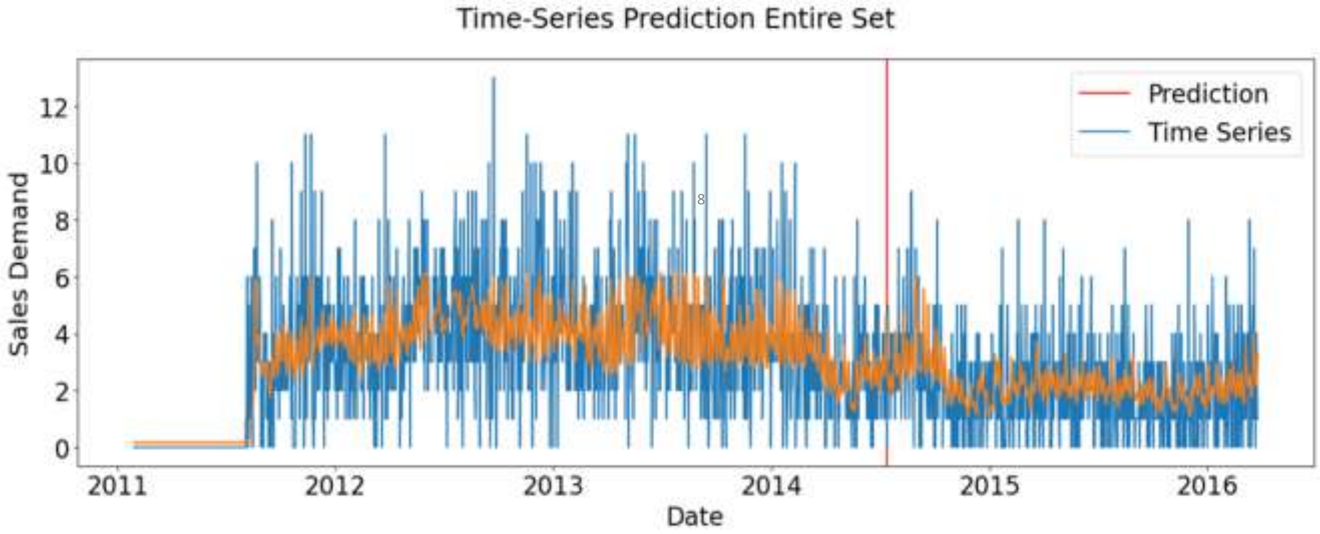
Hidden Size: 512

Number of Layers: 4

Optimizer: Adam with L2 Regularization

Learning Rate Scheduler: patience = 50, factor = 0.5

RMSSE - 0.76605



LIGHTGBM

Purpose: To understand LightGBM and explore its features

Approach Taken:

Reduced the dataset from 6 years to 2 years

Prediction carried out for random item selected from the dataset for next 28 days

Feature Engineering:

Merging relevant features of three dataset into one

Lag and rolling mean for last 7 days and 28 days

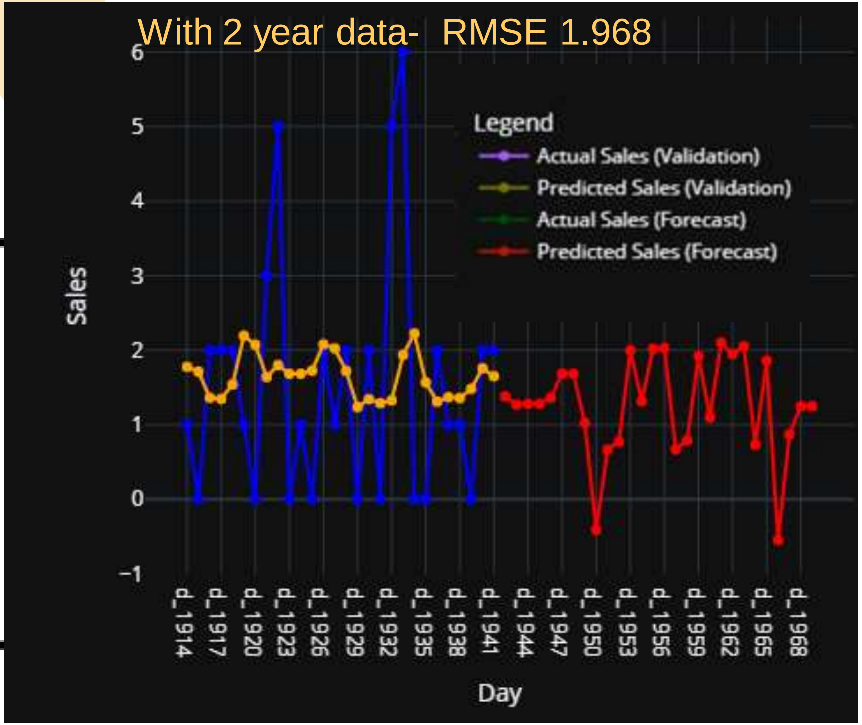
Evaluation Methodology:

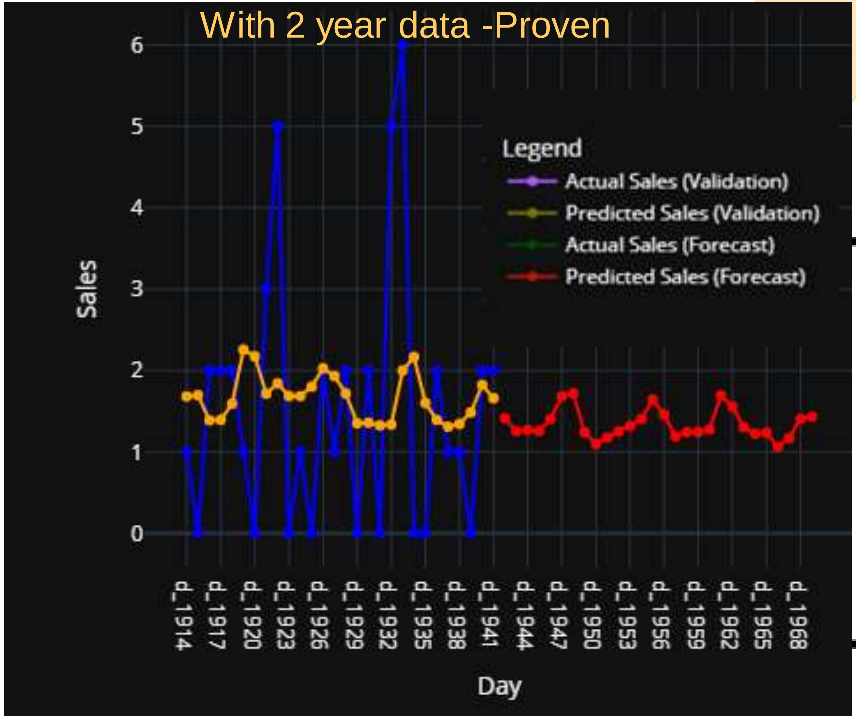
Generated forecast for random items and compare it with proven prediction from kaggle

Model Parameters:

```
lgb_params = {
    'objective': 'regression',
    'metric': 'rmse',
    'boosting_type': 'gbdt',
    'num_leaves': 31,
    'learning_rate': 0.1,
    'feature_fraction': 0.5,
    'seed' : 2000
}
```

LIGHTGBM-Results





Possible Reasons for difference in results:

1. Tweedie Regression objective function and other parameters considered for tuning
2. Recursive prediction

TEAM TASK ALLOTMENT

Sr.	Task	Executed By
1	Data Preprocessing	Prashik
2	EDA	Team
3	LSTM	Shoaib
4	Linear Regression	Prashik
5	LGBM	JatinKumar

Mr Ambuj Nayan(24M0003) has not participated in any project activity for unknown reasons.


Thank
You