Project Report – M5 Forecasting – Accuracy

Anchal GUPTA, Dinusara Sasindu GAMAGE NANAYAKKARA, Minji SEO, ZHAO Hang

Abstract

The M5 Forecasting Competition serves as a benchmark for retail forecasting methods, utilizing historical daily sales data from Wal-Mart stores. In this study, we perform exploratory data analysis (EDA), feature engineering, and model training to predict sales for the next 28 days. Our approach involves both time series models, such as Long Short-Term Memory (LSTM) networks and recurrent neural networks (RNN), as well as gradient boosting models like LightGBM and XGBoost. We employ various feature processing techniques tailored to each model's requirements and evaluate their performance based on Kaggle scores. Our findings suggest that the LightGBM model achieves the highest accuracy in predicting sales quantity.

1 Introduction

The M5 Forecasting Competition, the fifth iteration of the Makridakis time-series forecasting contest, offers a valuable benchmark for retail forecasting methodologies [1]. Utilizing historical sales data from Wal-Mart stores, this competition challenges participants to predict future sales trends. In this paper, we present our approach to forecasting sales for the next 28 days, involving exploratory data analysis (EDA), feature engineering, and model training. We explore both time series models, including Long Short-Term Memory (LSTM) networks[2] and recurrent neural networks (RNN) [3], and traditional learning models such as LightGBM [4] and XGBoost [5]. By leveraging various feature processing techniques tailored to each model, we aim to optimize predictive accuracy. Our study contributes to the field of retail forecasting by evaluating the performance of different modeling techniques in this competitive context.

2 DATA

The data in the form of grouped time series aggregated based on their type (category and department) and selling location (stores and states), with a total of 42,840 series .

In this project, we are given Walmart sales data and using this we are predicting the sales for next 28 days.

• Given data time period - Walmart's sales data from 29 Jan 2011 to 19 Jun 2016.

• The given forecasting task can be posed as a machine learning task and solved using ML algorithms. For this we will 1st need to re-frame data as supervised dataset and predict the units sold on a particular day using ML regression models.

• Our work mainly consists of the following 3 parts: exploratory data analysis (EDA), feature engineering, and model training

2.1 Data description inputs

The following four data sources were provided for the competition.

• sales_train_validation.csv: a file that contains the sales data from day 1 to day 1941. Each row represents sales data for a specific product in a particular store and state.

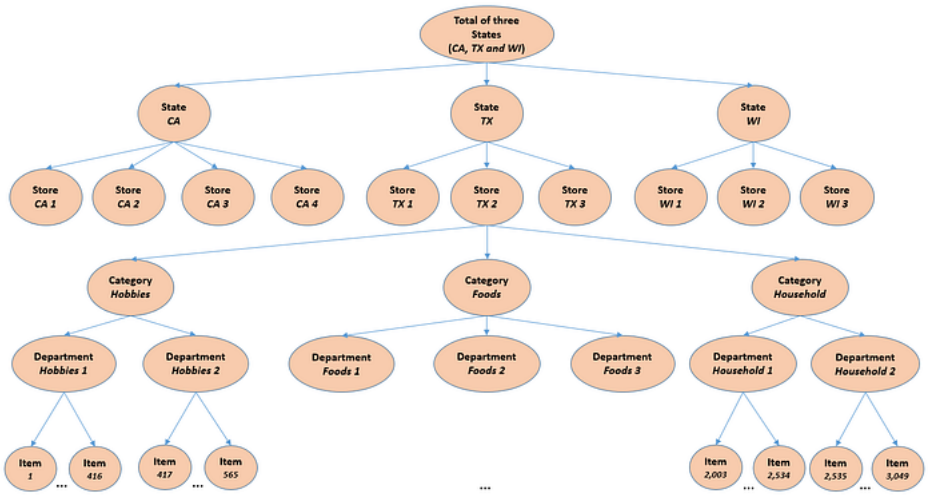

Figure 1: HOW DATA IS ORGANISED.

• sales_train_evaluation .csv: a file that contains the sales data from day 1 to day 1913. Data are stored in a similar format to sales_train_validation.csv.

• Calendar.csv: a file that contains information about the dates on which the products are sold. It also contains columns that represent the special events for a particular date.

• sell_prices.csv: a file that contains information about the price of the products sold per store and date.

2.2 EDA – Exploratory data analysis

From given data there are multiple ways to analyse the data to see which one would best fit our requirements. Some of the experiments were done using a representative data from sales_train_evaluation.csv file. This gave us an idea if that direction should be pursued further.

After trying out repeated patterns and methods, the time-based evaluation was clearly the best way to take forward. In sections below, we show our experiments and resultant data graphs.

cat_id	state_id	Count of cat_id
FOODS	CA	5748
	TX	4311
	WI	4311
FOODS Total		**14370**
HOBBIES	CA	2260
	TX	1695
	WI	1695
HOBBIES Total		**5650**
HOUSEHOLD	CA	4188
	TX	3141
	WI	3141
HOUSEHOLD Total		**10470**
(blank)	(blank)	
(blank) Total		
Grand Total		**30490**

Figure 2: SALES TRAIN EVALUATION.

Step 1. We analysed the data and created the charts below.

This gives us an idea on what is sales percentage across each of the products across categories or the sales percentage of the data across states which is selling them. The code file used to generate this data is attached here for reference -

Step 2. Sales across product category was analysed.

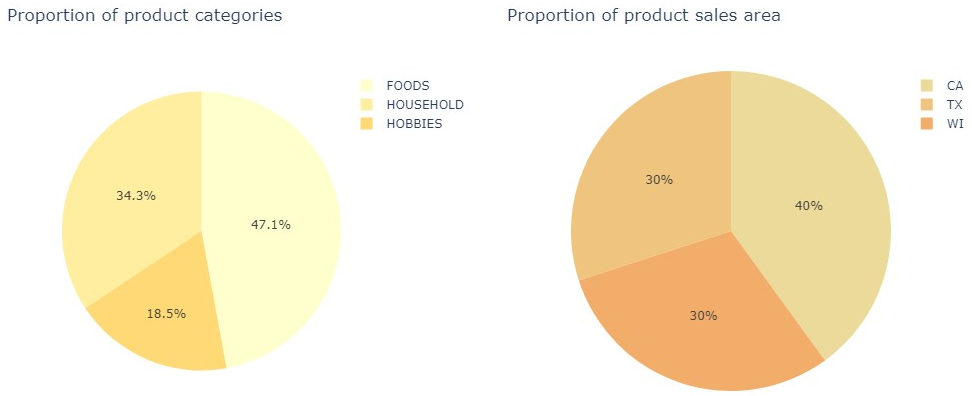

Figure 3: DATA ANALYSIS.

The below graph analyses the categories which are sold the most across the 2000 days. We can see that food item category sales have been consistently higher than others.

• California has recorded highest sales.

• FOODS category has highest sales at all stores followed by HOUSEHOLD and HOBBIES has minimum sales.

• At all stores FOODS_3 department in FOOD category has maximum sales followed by FOODS_2. FOODS_1 has minimum sales.

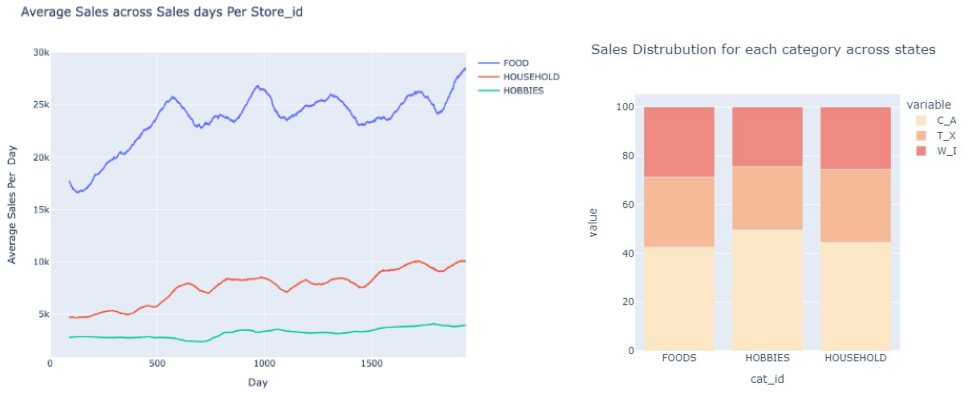

Figure 4: SALES OF PRODUCT CATEGORIES IN GIVEN TIME PERIOD.

Step 3. Sales at stores in given time was also analysed.

Figure 5 analyses how sales have been across the stores across the 2000 days. The store with code CA_3 has higher sales close to 6000 units, nearly 2000 units more than second highest sales making store.

• California sales have constantly increased whereas Texas sales are almost constant over the years. Wisconsin sales have two sharp increases in the given time period but overall has similar sales number as Texas.

Step 4. Data analysis for individual products sale in given time.

Figure 5: SALES ACROSS STORES IN GIVEN TIME PERIOD.

No event day	One event day	Two events day

Values	CA	TX	WI	Grand Total
Sum of d_80	9,955	7,218	5,216	22,389
Sum of d_50	14,509	8,203	7,887	30,599
Sum of d_501	14,676	10,051	10,347	35,074
Sum of d_1250	16,770	11,118	9,851	37,739

Values	CA	TX	WI	Grand Total
Sum of d_86	11,721	7,897	4,400	24,018
Sum of d_828	18,927	12,305	11,971	43,203
Sum of d_1178	18,353	12,553	8,028	38,934
Sum of d_1234	17,497	12,908	12,089	42,494

Values	CA	TX	WI	Grand Total
Sum of d_9	14,696	9,376	8,664	32,736
Sum of d_24	11,997	7,297	4,672	23,966
Sum of d_51	14,112	8,992	6,539	29,643
Sum of d_835	16,414	11,379	9,597	37,390

Figure 6: DATA ANALYSIS FOR INDIVIDUAL PRODUCTS VS STATE.

Figure 6 analyses how sales have been across the stores across the 2000 days for each of individual product. From CSV files using excel, if there were 2 events on a day the sales are higher than usual when there were no events. Figure 7 shows some examples of the sales data.

3 Feature processing

3.1 Time series model feature processing

In our approach to modelling sales data, we adopt a time series paradigm, considering the sequential nature of the data. Specifically, we utilize a retrospective window of 14 days' worth of sales data to forecast sales figures at a given time point. Additionally, we integrate event-related information sourced from the 'Calendar.csv' dataset by incorporating nine pertinent features into the feature set. These event features are appended to the sales data from the previous day, enhancing the predictive capacity of the model. Notably, this methodology is tailored for implementation within recurrent neural network (RNN) and long short-term memory (LSTM) architectures.

3.2 Gradient boosting model feature processing

Conversely, the preceding approach lacks utilization of crucial contextual attributes including item specifications, departmental categorizations, selling prices, and category classifications. To address this gap, we undertake feature engineering endeavours specifically tailored for traditional learning models such as LightGBM (LGBM) and XGBoost (XGBOOST). This entails the creation of new features derived from the temporal intervals surrounding each event occurrence. Each event is allocated its own feature column, wherein values denote the "strength" of proximity to the event, limited to a 30-day window. Moreover, we compute percentage differences in selling prices between consecutive weeks, generating additional columns reflecting price differentials. Furthermore, we compute rolling averages over 7 and 28-day intervals to encapsulate historical sales trends. Subsequently, this augmented feature set is employed to predict sales figures for a given day. Additionally, we adopt a stratified approach by segregating the data for each department (dept) and fitting separate models to each subset. This granular modeling strategy allows for more tailored predictions, capturing department-specific nuances and improving overall forecasting accuracy. Subsequently, these department-specific models are utilized to predict sales figures for respective departments on a given day.

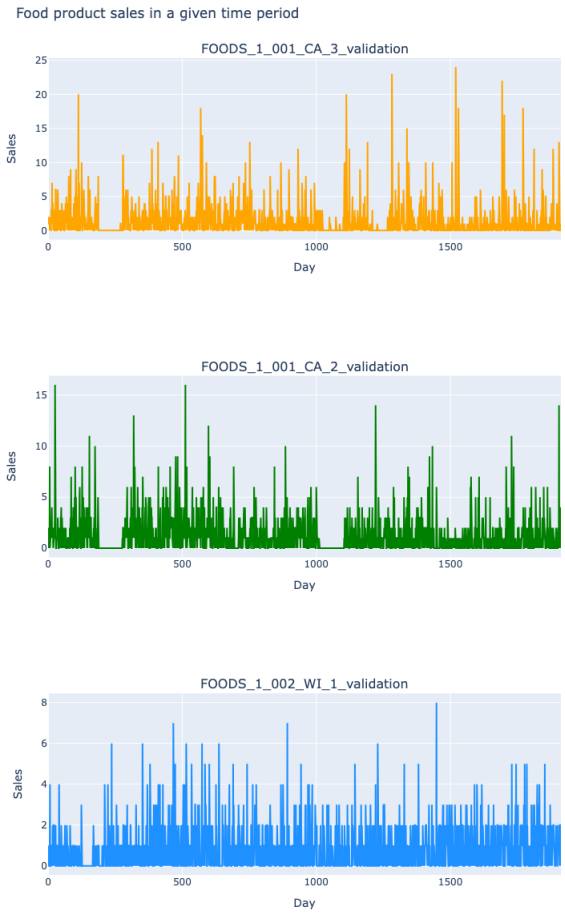

Figure 7: DATA ANALYSIS FOR INDIVIDUAL PRODUCTS VS STATE.

4 Model

We tried the LSTM, RNN, LGBM and XGBOOST models to predict the sales data.

4.1 LSTM model

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) that excel at capturing long-term dependencies in sequential data. They are commonly used for tasks like natural language modeling, speech recognition, and time-series analysis. After exploratory data analysis, based on the understanding of data we have designed new features. We have added features if event, week, and month. To prevent any bias from feature, all features were normalized.

Using the conclusion of EDA and decision to deploy time-period as basis, it was decided to train the LSTM model. The idea is to use the point(epoch)where training loss and validation loss is at minimum.

Step 1 – Use final_data_scaled array

Step 2 – we used 90% of data as training data and 10

Step 3 – Based on that, calculated mean squared error (MSE) as loss function.

Step 4 – From the graph for visual confirmation and code for exact point, we can what point to use where validation and training loss is minimum.

Step 5 – Use that point as point where model was trained.

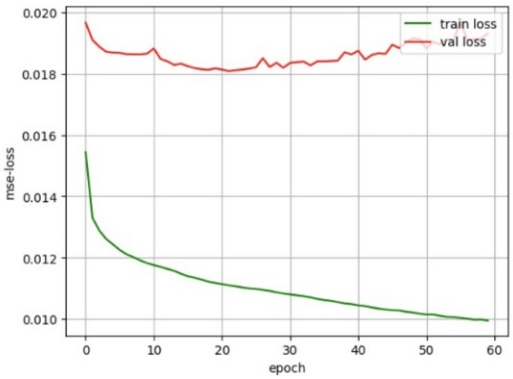

Figure 8: MSE LOSS FOR TRAINING AND VALIDATION DATA.

Step 6 - For the time selected in Step 5 for validation loss of 0.0181, we train our model for next 28 days as per the original problem.

4.2 Recurrent neural network model

Recurrent Neural Networks is a deep learning model that is effective in handling sequential data by identifying long term dependencies The recurrent neural network (RNN) model included three recurrent layers for capturing long term dependencies and the output of the RNN layers is finally passed to Dense layer for making predictions. In between the RNN layers, Dropout layers are added to prevent overfitting. The model was trained with the following parameters.

Table 1: Table of parameters and corresponding values for RNN model training

Parameter	Value
Optimiser	Adam (Chandriah, n.d.)
Number of epochs	60
Loss function	Mean squared error

Figure 9 plots the training and validation loss respect to the number of epochs.

4.3 Light GBM model

Light Gradient Boosting Machine is a gradient boosting method designed based on decision tress to improve computational efficiency and prediction accuracy.

Since we wanted to explore how the prediction (number of sales) was dependent of variables that are counts of per unit time we decided to use LGBM (LightGBM, n.d.) Poisson regression model. The model was trained with the following parameters:

Figure 10 is the visualization of decision tree visualization in the model. Figure 11 is the LGBM feature importance.

Table 2: Table of parameters and corresponding values for LGBM model training

Parameter	Value
Learning rate	0.09
Number of epochs	2000
Loss function	Negative Log Likelihood

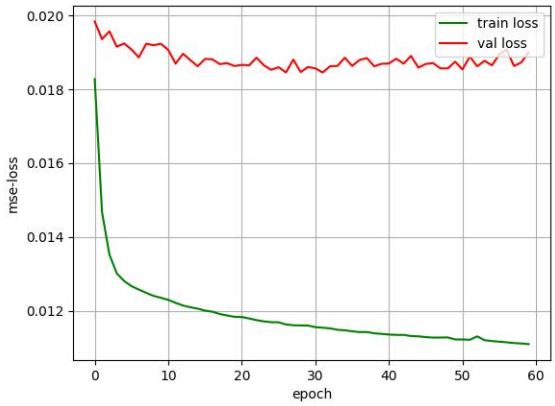

Figure 9: RNN MODEL.

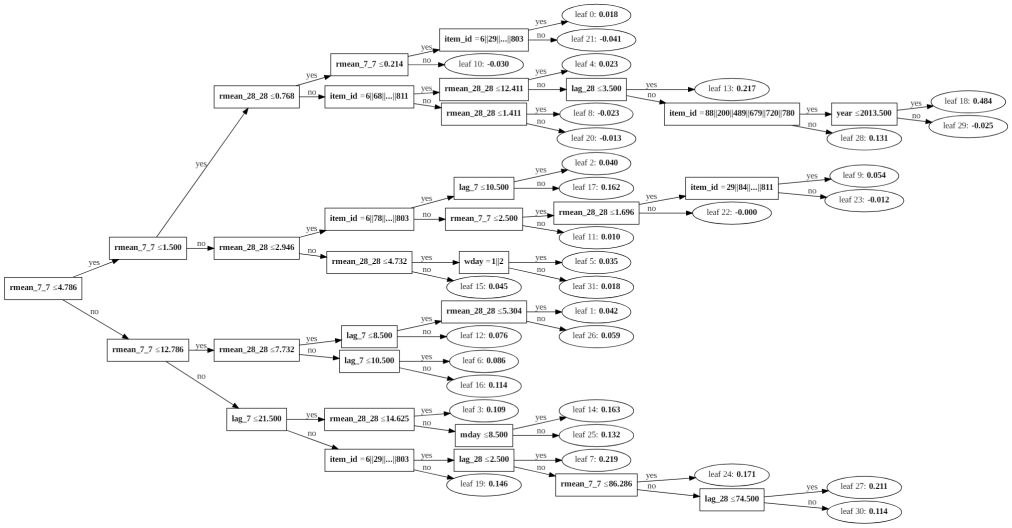

Figure 10: LGBM VISUALIZATION.

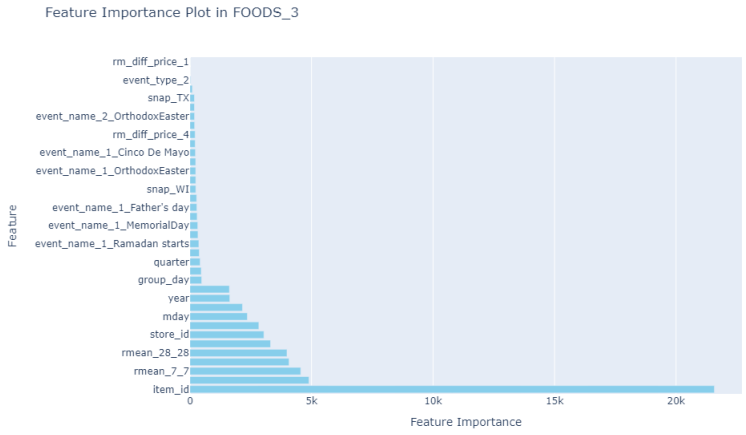

Figure 11: LGBM FEATURE IMPORTANCE.

4.4 XGBoost model

Extreme Gradient Boosting is a grading boosting method implemented based on the gradient descent optimization technique.

We implemented a Poisson regression model with the following parameters:

Table 3: Table of parameters and corresponding values for XGB model training

Parameter	Value
Learning rate	0.09
Number of epochs	2000
Loss function	Negative Log Likelihood

After training the XGBoost (XGBoost Documentation, n.d.) model we also plotted a feature importance to identify which features has the most impact in predicting future sales quantity. From the below plot it can be deduced in predicting sales quantity for future data is dependent on the previous month/week sales quantity.

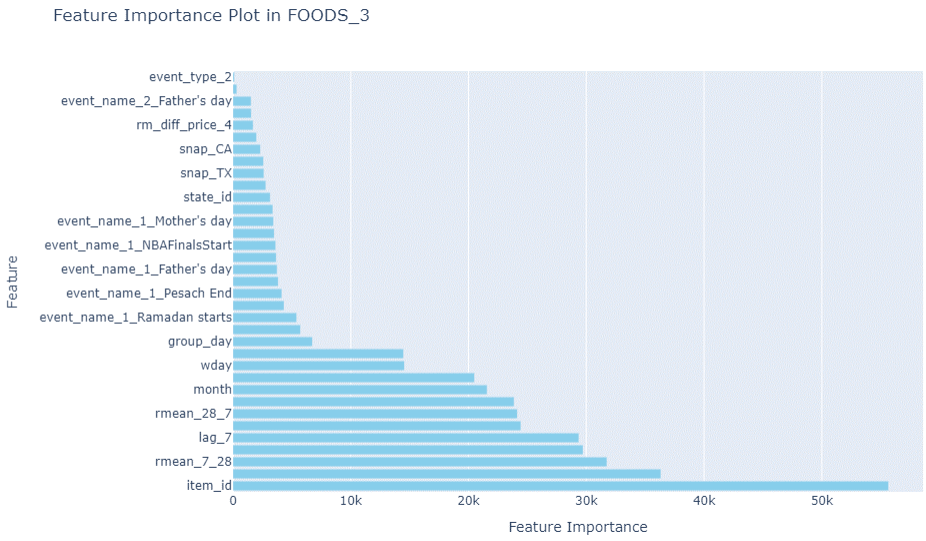

Figure 12: PLOT TO DEMONSTRATE FUTURE DATA IS DEPENDENT ON PAST DATA.

4.5 Model performacne

The below table shows the Kaggle private and public scores for our models. Each score represents the mean error from the test data per model. We can clearly see that LGBM model has lowest public and private score which makes it our choice.

5 Conclusion

In conclusion, our study underscores the effectiveness of various modeling techniques in forecasting retail sales. Through exploratory data analysis (EDA), feature engineering, and model training, we have thoroughly assessed the performance of time series models like LSTM and RNN, as well as traditional learning models such as LightGBM and XGBoost. Our findings indicate that the LGBM model outperforms other models, achieving the highest accuracy in predicting sales quantity. Moreover, our analysis reveals that deep learning models like LSTM and RNN, while capable of capturing complex temporal patterns, may face challenges in accurately predicting integer-valued

Table 4: Model performance summary

Model	Private score	Public score
LSTM	1.36598	1.4484
RNN	1.21168	1.25751
XGBOOST	0.58911	075612
LGBM	0.58542	0.74072

Figure 13: MODEL PRIVATE AND PUBLIC SCORE.

sales due to their inherent tendency to produce continuous outputs. Despite this limitation, leveraging a combination of deep learning and traditional machine learning techniques can provide valuable insights into retail sales forecasting. This study emphasizes the importance of model selection and feature processing in optimizing predictive performance and highlights the need for further research to refine and integrate diverse modeling approaches for enhanced retail sales prediction.

6 Authorship Contribution Statement

• Anchal GUPTA – Design of the overall project flow and writing of section 2

• Dinusara Sasindu GAMAGE NANAYAKKARA – Data analysis of LGBM model and writing of section 4

• ZHAO Hang – Data analysis of EDA, RNN, LSTM, XGBoost, LGBM model, writing of section 1,3,5 and LaTeX formatting

• Minji SEO – Analysis check and slides/presentations

References

[1] Addison Howard, i. S. (2020). *M5 Forecasting - Accuracy*. Retrieved from Kaggle: `https://kaggle.com/competitions/m5-forecasting-accuracy`

[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780.

[3] Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179-211.

[4] *Aziz, R. M., Baluch, M. F., Patel, S., & Ganie, A. H. (2022). LGBM: a machine learning approach for Ethereum fraud detection. International Journal of Information Technology, 14*(7), 3321-3331.

[5] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).