Segmentation and Forecasting for Premier E-Tailer

Yuntong Lin [1], Abhishek Talwar [2], Ming-Jen Yeh [3], Daniel L. Whitenack [4]

Purdue University Krannert School of Management

[1] lin1053@purdue.edu; [2] talwar2@purdue.edu; [3] yeh50@purdue.edu; [4] dwhitena@purdue.edu

Abstract

Through an empirical study of a premier e-tailer with a consumer base of students from **over 1,400 universities** and with a product line of **over 20,000 products**, this project assesses **how advanced analytics can be applied to forecast revenue and improve performance**. This project focuses on scrutinizing revenue streams through different business lines from 2009-18 and forecasting revenue for 2019 while highlighting peak season time-frames for different lines of business and evaluating how these insights can be translated to tangible actions.

Introduction

E-commerce is growing at a pace faster than any retail sector, and this has resulted in a quest for **competitive advantage** which in turn demands greater **scrutiny of revenue flows**.

Through **engagement with a premier e-tailer** dominating the university merchandise market with a target consumer base of students **from over 1,400 schools** and engaged in retail of **over 20,000 products** through **6 lines of business**, this project focuses on **drawing actionable insights** by scrutinizing their revenue streams through **different business lines** from 2009-18 and **forecasting revenue for 2019**.

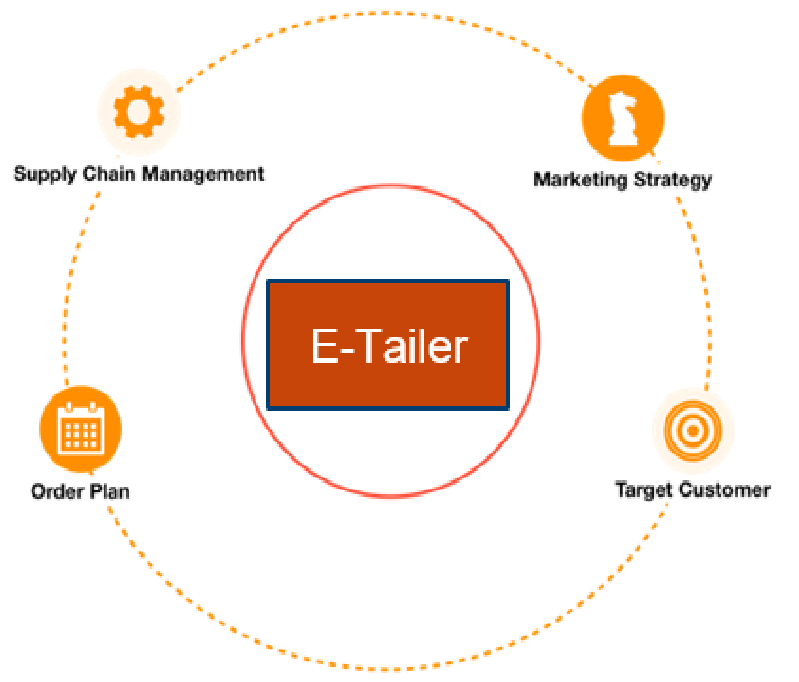


Figure 1. Ecommerce sales cycle

By forecasting the expected revenue of 2019, this project can help the firm to:

- Foresee the upcoming business variations
- Transform the business insight into tangible actions to better plan for the next business cycle.

Literature Review

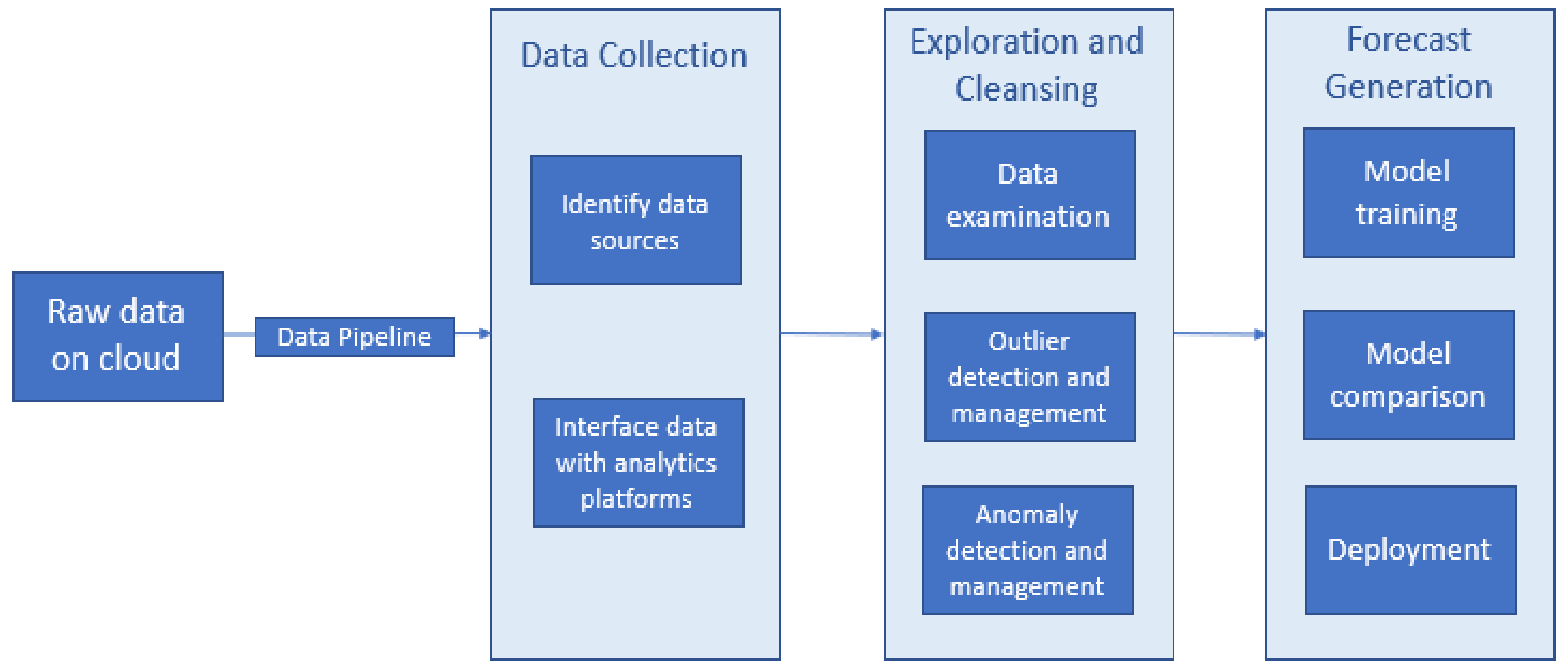
As business world becomes more and more complex, recent work on sales prediction used different models to effectively predict the customer need. Below is the how previous study achieved in understanding of the complex dynamical patterns in the linear regression, time series, and machine learning methodology.

Study	ARIMA	Lasso	Tree Based Classifier	LSTM
(White and Ariguzo, 2011)	✓			
(Arunraj and Ahrens, 2015)	✓			
(Pavlyshenko, 2016)	✓	✓		
(Pavlyshenko, 2019)	✓	✓	✓ (Random Forest)	
Our Study	✓	✓	✓ (XGBoost)	✓

Table 1. Literature review summary by method used

Since the approach of conducting time series analysis with machine learning method can add more interpretable insight (Fulcher, 2018), we also took Long-Short Term Memory (LSTM) recurrent neural network in our analysis.

Methodology



Data Collection
- Identify data sources
- Interface data with analytics platforms

Raw data on cloud → Data Pipeline → Data Collection → Exploration and Cleansing → Forecast Generation

Exploration and Cleansing
- Data examination
- Outlier detection and management
- Anomaly detection and management

Forecast Generation
- Model training
- Model comparison
- Deployment

Figure 2. Study Design

Data

The dataset consists of student databases from target universities along with information pertaining to purchase orders (product ordered, quantity ordered, net revenue order) in the time frame 2009-19.

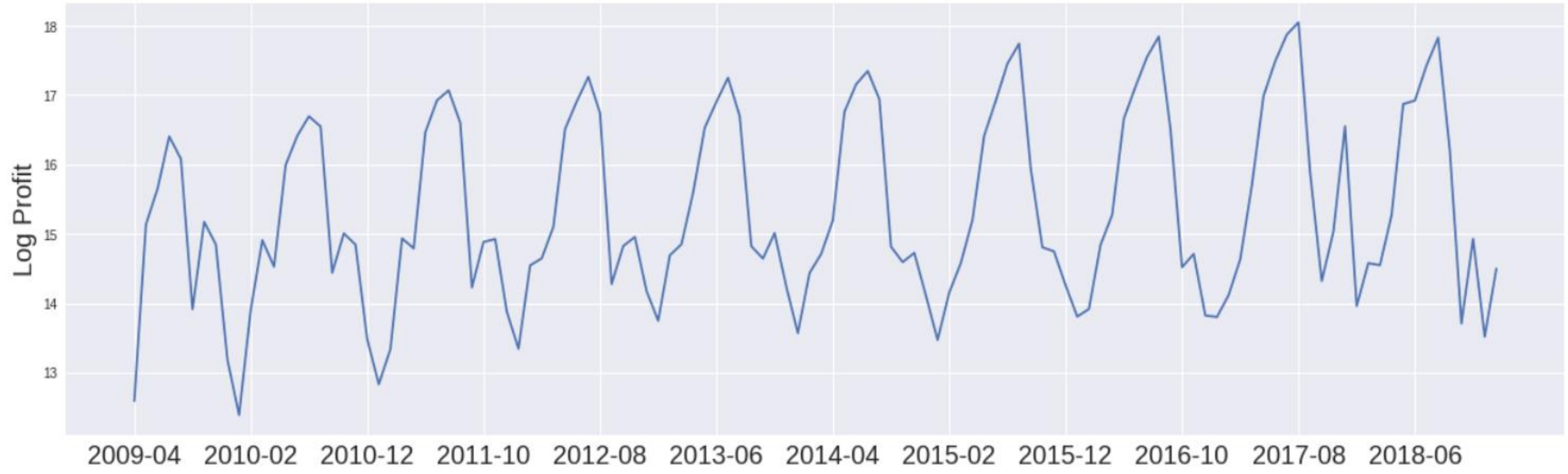


Figure 3. Profit trends

Data Collection

The purchase orders data for the e-tailer was hosted on a cloud based platform and was pulled in through for analysis on cloud based tools.

Exploration & Cleansing

The dataset contains some anomalies to be processed. Overall, key columns with products information along with purchase records have been retained and organized.

Forecast Generation

- Model training:
  Different models (ARIMA, LSTM, Lasso, XGBoost) were trained on purchase records (aggregated on LOB and Month-Year) from 2015-2017.
- Model Comparison:
  These models were then used to predict LOB revenue for 2018 and these predictions were compared with the actual observations for 2018. Through this comparison using MAPE (Mean Absolute Percentage Error) as a measure, ARIMA was identified as the optimal model with the lowest MAPE.
- Deployment:
  Finally, once the optimal model was identified, the model was re-trained on the dataset from 2015-18 and forecasts 2019 were generated.

Models

Model Selection

We have two type of models – time series models ARIMA and LSTM and regression models – Lasso and XGBoost to simulate time serious by regression.

Model Evaluation / Statistical & Business Performance Measures

We used MAPE as our performance measure to get a rough picture of how our models can perform. We choose this method because MAPE can measure the error in percentage and applies absolute value to eliminate the impact of negative values. Also, it is more intuitive to understand than RMSE that is used to be a usual measure in LSTM.

Results

MAPE Comparison

Below is the comparison of MAPEs of all four methods. As we can see, ARIMA model performs the best. Thus, we will use ARIMA(1,12,1) to forecast future profits.

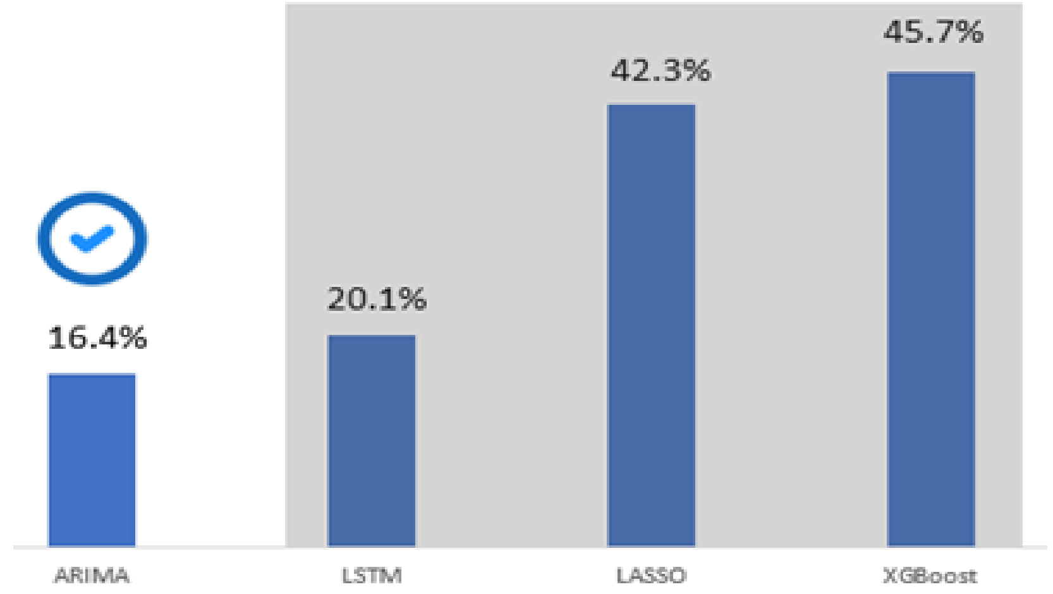


Figure 4. MAPE Comparison

Business Forecast

Below are the predicted profits for the 6 lines of business:

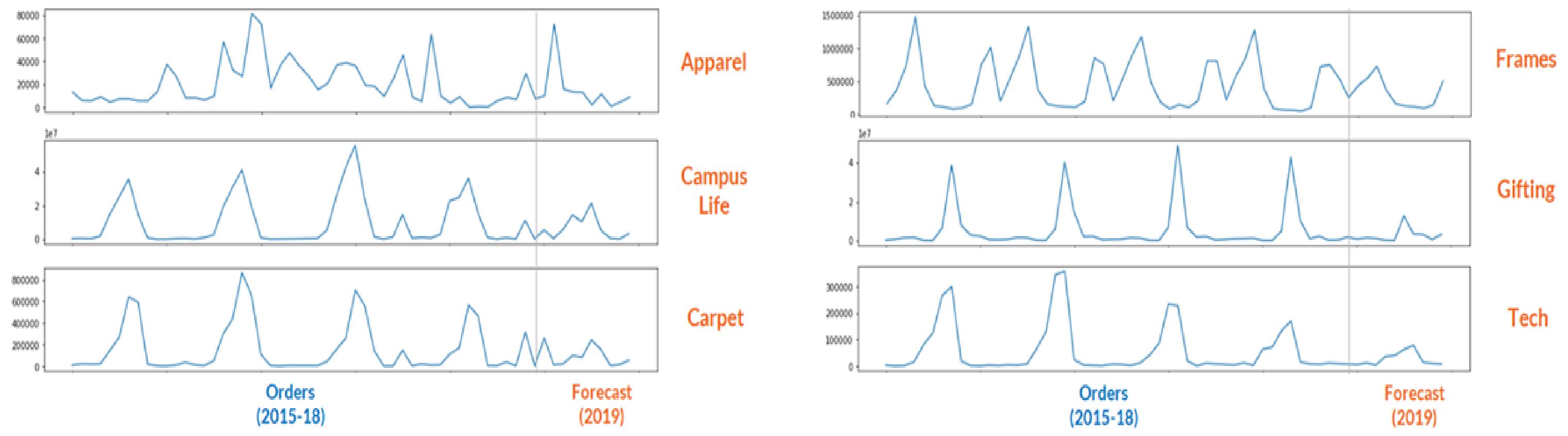


Figure 5. Revenue forecast (2019)

Conclusions

Through analysis of historic sales data of a premier E-tailer's purchase records from 2009-18, revenue for 2019 was forecasted by using an ARIMA model. Insights generated through this forecast highlight peak seasons where the E-tailer can augment their supply-chain and marketing efforts to make the most of the high demand. The model deployed are applicable in the broader e-retail space and the insights generated, while unique to different organizations, help provide a platform of driving business growth in future.

Acknowledgements

We thank Professor Daniel Whitenack for constant guidance on this project.