CECS 551 Advanced Artificial Intelligence

Final Project

Fall 2022

Artificial Intelligence (AI) approach in Retail Market Analysis and Growth

1. Introduction

The final project entails analyzing inventory data of two dataset of around more than 30 stores of an international retail business. It is designed to implement two-week sprints of the scrum process, mimicking a real tech company software development and machine learning work environment. The purpose of the analysis is to use the inventory data to improve sales, resulting in a more efficient operation.

- **CECS551 dataset 01**: The task is to predict the department-wide sales for each store.
- **CECS551 dataset 02**: The goal is to predict the unit sales of each product for the next 10 days from 10 different stores across the two states, i.e., STATE1 and STATE2.

You will be scored for **100 points** in total spanning across **three sprints**. Please see the timetable below for more details. It is suggested that one person plays the role of the scrum master to coordinate the communications between team members and ensure on-time delivery at the end of each 2-week sprint.

1.1 Sprint schedule

Table 1: Project schedule over 3 Sprints

Sprint	Date	Deliverable	Points	Feedback
Sprint 01	10/17 – 10/30	Chapter 01 - Report, Presentation, Code	30	email
Sprint 02	10/31 – 11/13	Chapter 02 - Report, Presentation, Code	40	Zoom
Sprint 03	11/14 – 11/27	Chapter 03 - Report, Presentation, Code	30	Zoom

1.2 Dataset

There are two datasets for the project, and you can download the dataset here (CSULB SSO login is required for accessing the dataset).

Report: Each team member should document their contributions in each report.

Presentation: Guidelines for presentation will be shared.

2. Retail analysis using Artificial Intelligence approach

2.1 Sprint 1: Data visualization

Tableau; Python

As a data scientist, you must present the information to an upper higher-level management of an organization who want to get the snapshot of the data and understand the current business from the give dataset. You are advised to use Tableau for the visualization for the two datasets. The goal is to create a Tableau dashboard and publish the results. There can be certain cases, where you might want to use visualization using Python libraries. Please use the right data visualization method for specific problem statement.

2.1.1 Analyze the dataset for CECS551 dataset 01

1. Visualize the weekly and monthly sales pattern across top 35% of the stores based on the “type” of the products, department, and the “size” of the store. Identify the best department and product “type” across the first ten stores.

2. Investigate the relationship between weekly sales over CPI, unemployment, and holiday for the first 10 stores. You can explore the what-if scenarios while writing the report.

Choose the right chart type, for example, box-plot, histogram, scatter-plot, pie-chart, etc.

3. What is the impact of various types of discounts, for example, discount promotional, dis- count clearance, discount damaged good, discount competitive and discount employee on the overall sales. Which type of discount is helpful in increasing the sales? Does the observed behavior hold true for all the stores? You may consider using the data from 40% of the stores where the sales are low.

4. Identify the “type” of products which are highly impacted by external factors like “temperature”, “gas price”, and “holiday”. Is there any correlation between overall sales and holiday?

2.1.2 Analyze the dataset for CECS551 dataset 02

1. Visualize the daily, monthly, and total sales in STATE1 and STATE2. Identity the department with highest and lowest sales across all stores. You may make a comparison of “State” with “Item Categories”.

2. Determine the product sales and availability over time in each store, and average price on each category.

3. Visualize the monthly and total sales for each store and each category. Identify the most sold products in among each state in each department.

4. Investigate the average sales of highly sold product for given price of each month and on weekdays, and average sales on event types, for example, cultural, national, etc.

2.2 Sprint 2: Machine Learning model

Designing a machine learning model is an iterative process where we start with a baseline model and improve its performance by observing the performance metrics and the end goals remains optimizing the model.

In sprint 2 you will work on two different dataset and solve the retail problem statement using machine learning. You must share the though process and key findings while handling each problem statement. Always communicate the performance metrics of each model as supporting evidence for your findings.

2.2.1 CECS551 dataset 01

1. Identify the key variables for the model using correlation plots, heatmaps, histograms, feature importance (SHAP). explainability

2. Forecast the weekly sales across first ten stores and use the same model to make predictions for store_11_35. Is modeling averages by department or store more likely to produce superior results?

3. Design a regression model to forecast the weekly sales using the given features. Begin with Linear regression and observe the accuracy of the model. Regression

4. As expected, linear regression is the not optimal algorithm for the problem statement. You are encouraged to explore other regression algorithms to design, evaluate, and improve model accuracy (try statistical approach and deep learning models). Support your finding by validating the model accuracy across various stores. GBM, XGBoost, RF)

5. Estimate the impact of external variable such as gas price, holidays, unemployment, and temperature on overall sales across the stores. Do you observe any relationship between Weekly Sales and Unemployment?

6. In stores.csv, we have a feature “type”, i.e., three types of stores - A, B, and C. Now, based on the sales volume can we predict the store type? You may treat this as multi-label classification problem. You can use any other classification algorithm but compare the performance metrics of each classifier being used.

Classification Hyper-parameter tuning

Reason for selecting the algorithm Deep Learning model and Statistical model

2.2.2 CECS551 dataset 02

The task is to design a machine learning model to make accurate predictions for product sales for next 5 and 10 days in advance and compare the model performance.

1. You may want to experiment with various statistical techniques and deep learning models to find an optimal approach. For example, compare LSTM (along with hyper-parameter tuning), LightGBM for the time series forecasting and evaluate the performance metrics of the winning algorithm.

Time series

trend, seasonality

2. Experiment if external variables are important for improving the forecasting accuracy of time series methods.

ARIMA - baseline

n-step ahead forecasting

3. Identify the winning algorithm and reason out if the same algorithm will perform optimal on individual stores.

4. Cluster the stores based on similar products and price (estimate the optimal number of clusters). You may club together the sales across stores for each category. You may begin with k-means clustering as a baseline and compare with other clustering algorithms.

Category	Clusters
EDIBLE	x
ENTERTAINMENT	y
HOUSE	z

Clustering

Comparison and performance metrics

Note: Extra credit will be considered for team's effort on model improvement. Improving the machine learning models is where a data scientists will shine in their career and ahead of the game from others.

2.3 Sprint 3: Model deployment and business recommendation

The work of a Data Scientist doesn't end by designing an accurate model with the historical data. The model should also be deployed for the end user. Moreover, once the model is deployed the actual challenges of monitoring the performance degradation of the model comes into the picture which should be handled as part of end-to-end implementation of data science life cycle. Sprint 3 deals with addressing these problems.

1. Deploy the winning model on Heroku for at least one dataset. You can also publish your model on Explainerdashboard.

2. Once the model is deployed in production, it's not guaranteed that the model will perform as expect, and sometime the performance deviates from the expected. The task is to estimate model performance estimation and monitoring post deployment. You may want to explore nannyML for this problem statement. (Please do not use any proprietary codes)

IMPORTANT

3. As a business analyst, provide at least three recommendations to increase the sales across each department and marketing strategies. Discuss the impact of the initiatives separately and use relevant graphs to support the recommendation.

4. Identify and report the challenges you might have faced in sprint 2 and sprint 3.

3. Coding and report requirements

1. Data should be saved on Google Drive and loaded to Google Colab.

2. Codes should be developed professionally with proper documentation of notes, assumptions and variable definitions for your teammates and others to easily understand and follow.

3. The report should be a Google Doc file prepared in collaboration between team members.

4. Follow the best practices for data visualization. Use relevant graphs for specific problem.

6. Tell your story, provide insights, and organize your charts to support your thoughts.

7. Provide an executive summary at the beginning of each chapter and final conclusions at the end.

4. Final submission

Publish the project report, presentation, and source code on GitHub. Please add your team members as collaborators.

1. Link to the code on Colab and exported python code

2. Link to the Google Doc report and the exported file in MS Word format (.docx)

3. PDF of the final Report

4. Final Presentation

5. Link to the published Tableau dashboard

5. Data description

CECS551_dataset_01

store 01-10.csv

store - The number of stores date - MMDDYYY format

temperature - Temperature in Fahrenheit gas price - Price per gallon in $ discount promotional - discounts discount clearance - discounts

discount damaged good - discounts discount competitive - discounts discount employee - discounts

CPI - The Consumer Price Index (CPI) Unemployment - Unemployment rate in the region where store is present

IsHoliday - Yes or No

stores.csv

store - The number of stores

type - Stores segregated into three types, i.e.,

A, B, and C

size - Size of the store

test.csv

store - The number of stores dept - Department ID

date - MM/DD/YYYY IsHoliday - Yes/No

train.csv

store - The number of stores dept - Department ID

date - MM/DD/YYYY weekly sales - Sales per week IsHoliday - Yes/No

CECS551_dataset_02

calender.csv

date – date

weekday – categorical

wday – weekdays in numeric month – month in numeric year – year in numeric

d – each day assigned in sequential order

event name 1, event name 2 – the name of events

event type 1, event type 2 – the type of events

snap STATE1 and STATE2 - snap is a nutritional program for low-income families.

data_test.csv and **data_train.csv**

id – product id

item id – items

dept id – department

cat id – category

store id – store id

state id – state id

d 1 - d 1941 – day 1 to day 1941

price.csv

store id - store id

item id – item id

sell price – selling price

■