CECS 551 Advanced Artificial Intelligence

Final Project

Fall 2022

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

# 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 for 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 various states.

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 sprint.

## Sprint schedule

**Table 1:** Project schedule over 3 Sprints

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sprint** | **Date** | **Deliverable** | **Points** | **Feedback** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sprint 01 | 10/25 – 11/04 | Chapter 01 - Report, Presentation, Code | 30 | email |
| Sprint 02 | 11/05 – 11**/**18 | Chapter 02 - Report, Presentation, Code | 40 | Zoom |
| Sprint 03 | 11/19 – 12/03 | Chapter 03 - Report, Presentation, Code | 30 | Zoom |

**Sprint Review on Zoom**

* Sprint 02: 11/12 (morning)
* Sprint 03: 12/03

**Final submission**: 12/04/2022.

**Final presentation**: 12/06 and 12/07 [7:30 PM – 9:30 PM] over **Zoom**.

## Dataset

There are two datasets for the project, and you can download the dataset [here](https://csulb-my.sharepoint.com/:f:/g/personal/rahuldeo_vishwakarma01_student_csulb_edu/EntULpvD461NopdRdApcKZMBKdKahWoBCZMvL_ZAjok1wQ) (CSULB SSO login is required for accessing the [dataset](https://csulb-my.sharepoint.com/:f:/g/personal/rahuldeo_vishwakarma01_student_csulb_edu/EntULpvD461NopdRdApcKZMBKdKahWoBCZMvL_ZAjok1wQ)).

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

# Retail analysis using Artificial Intelligence approach

## 2.1 Sprint 1: Data visualization

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](https://www.tableau.com/products/desktop) for the visualization for the second dataset. The goal is to create a Tableau dashboard and publish the results.

### Analyze the dataset for CECS551\_dataset\_01

Perform exploratory data analysis using **Python** for dataset\_01. Please use the right data visualization method for specific problem statement (choose the right chart type, for example, boxplot, histogram, scatterplot, pie-chart, etc.).

1. Identify the key variables for the model using correlation plots, heatmaps, histograms, feature importance (SHAP).
2. For the first 10 stores visualize the weekly and monthly sales patterns for top 35% of the product sales and department sales.
   1. Identify the best department and product “type” across the first ten stores.
3. Investigate the relationship between weekly sales over CPI and unemployment for the first 10 stores. You can explore the what-if scenarios while writing the report.
4. Investigate the impact of various types of discounts, for example, discount promotional, discount clearance, discount damaged good, discount competitive and discount employee on the overall sales.
   1. Which type of discount is helpful in increasing the sales? Consider top 30% of the best performing stores (sales per 1000 square feet).
   2. Does the observed behavior hold true for all the stores? Consider bottom 30% of the least performing store (sales per 1000 square feet).
5. Identify the “type” of products which are highly impacted by external factors: “temperature”, “gas price”, and “holiday”. Is there any correlation between overall sales and holiday?

### Analyze the dataset for CECS551\_dataset\_02

The dataset represents the unit sales (in number/quantity) of various products sold in the USA, organized in the form of grouped time series. More specifically, the dataset involves the unit sales of around 3000 products, classified in 3 product categories (Hobbies, Foods, and Household), and 7 product departments in which the above-mentioned categories are disaggregated.

The products are sold across 10 stores, located in 3 States (CA, TX, and WI). In this respect, the most disaggregated data, i.e., product- store unit sales, can be grouped based on either location (store and state) or product-related information (department and category).

Diagram

Description automatically generated

1. Use Tableau to visualize the dataset\_02.

<https://www.tableau.com/products/desktop>

1. Publish the Tableau dashboard on public server.

For example, <https://public.tableau.com/app/profile/joslininsight/viz/AONLLongitudinalL4NursingLeadershipInsightStudy/AONLLongitudinalL4NursingLeadershipInsightStudy>



Example of dashboard (different dataset)

Graphical user interface, chart, application, scatter chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

## 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.

* Dataset\_01: Regression, Classification
* Dataset\_02: Time Series

You must share the though process and key findings while handling each problem statement. Communicate the performance metrics of each model as supporting evidence for your findings.

### 2.2.1 CECS551\_dataset\_01

The data consists of 45 stores including store information and monthly sales. The objective is to predict sales of store in a week. As in dataset, size and time related data are given as feature, so analyze if sales are impacted by external factors, for example, how inclusion of holidays in a week soars the sales in store?

1. Design a prediction model to forecast the weekly sales across first ten stores and use the same model to make predictions for store\_11\_35. There are external variables such as gas price, holidays, unemployment, and temperature for the given dataset. Evaluate the impact of these external variables on the accuracy of the model (do they help improve the accuracy?). Plot the relevant graphs.
   1. Begin with Linear regression model to forecast the weekly sales using the given features.

***Hints***: Feature selection, feature engineering (new features, if any), PCA.

* 1. Create the following machine learning models: ARIMA, Ridge Regression and Boosting to predict sales.
  2. Communicate the model performance metrics and tabulate the comparison in report. Support your finding by validating the model accuracy across various stores (first 10 stores and store\_11\_35).

|  |  |  |
| --- | --- | --- |
|  | First 10 stores (combined data) | store\_11\_35 |
| Algorithm 1 | Performance metrics | Performance metrics |
| Algorithm 2 | Performance metrics | Performance metrics |
| Algorithm n | Performance metrics | Performance metrics |

1. In stores.csv, we have a feature “type”, i.e., three types of stores – **A** (Super center), **B** (Discount center), and **C** (Neighborhood markets). Now, based on the given features in the dataset, can we predict the store type? Please consider only first 10 stores for this problem statement.
   1. Consider the problem statement as multi-label classification problem. Use the below classification algorithms and perform hyper-parameter tuning for the Deep Learning models.

* Ensemble models (3 statistical methods)
* Recurrent neural network (RNN)
* Convolutional Neural Networks (CNN)
  1. Plot the relevant graphs and tabulate the performance metrics (ROC, AUC, Precision-Recall, confusion matrix, F1 score).

### 2.2.2 CECS551\_dataset\_02

**Task 1**: Design a machine learning model to make accurate predictions for product sales for next 10 days in advance (the data set includes daily unit sales per product) and compare the performance of different machine learning algorithms.

**Background**: The dataset also involves external variables, for exmaple, calendar-related information and selling prices. Thus, apart from the past unit sales of the products and the corresponding timestamps (e.g., date, weekday, week number, month, and year), there is also information available about: special events and holidays (e.g., Super Bowl, Valentine’s Day, and Orthodox Easter), organized into four classes, namely Sporting, Cultural, National, and Religious.

Selling prices, provided on a week-store level (average across seven days). If not available, this means that the product was not sold during the week examined. Although prices are constant on a weekly basis, they may change with time.

1. Perform data preprocessing and exploratory data analysis.

***Hints:*** Does seasonality and trend exist in the dataset?

Drop any column?

Extra credit: Perform down-casting (shrink dataset size) <https://pypi.org/project/pandas-downcast/>

1. Feature engineering: create two new features using the information provided in Table 1.

a) weather data

b) median income

You need to consider the weather information of each store location in particular state and use it as a feature to build the forecasting model. Use the below information to retrieve the weather data using the corresponding Zip code of each location.

|  |  |  |
| --- | --- | --- |
| **Code** | **Location** | **ZIP Code** |
| CA1 | Long Beach | 90804 |
| CA2 | Los Gato | 95032 |
| CA3 | San Diego | [92107](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/ZIP:92107/detail) |
| CA4 | Fremont | [94536](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/ZIP:94536/detail) |
| TX1 | [Rockwall](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/ZIP:75032/detail) | 75032 |
| TX2 | [Frisco](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/ZIP:75033/detail) | 75033 |
| TX3 | [Sachse](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/ZIP:75048/detail) | 75048 |
| WI1 | [Watertown](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/ZIP:53094/detail) | 53094 |
| WI2 | [Mequon](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/ZIP:53092/detail) | 53092 |
| WI3 | [Theresa](https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/ZIP:53091/detail) | 53091 |

***Hint:***

Weather data: <https://www.climate.gov/maps-data/dataset/past-weather-zip-code-data-table>

Median income: <https://data.census.gov/cedsci/table?q=ZCTA5%2090804%20Income%20and%20Poverty&tid=ACSST5Y2020.S1903>

Graphical user interface, text, application, email

Description automatically generated

1. Use the below machine learning algorithms to model n-step ahead forecasting (n = 10).

Note: First create model without using any external features, and then create model with the external features.

* Begin with ARIMA and compare the RMSE values for each category.
* Long short-term memory (LSTM) – Perform hyper-parameter to improve the model.
* Plot the relevant graphs and tabulate the performance metrics of each model.

**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 and provide some decision making for business refinement. Sprint 3 deals with addressing these problems.

1. Deploy the winning model on [Heroku](https://www.heroku.com/) for dataset\_02 (ARIMA)

3. The logistic performance is impacted by product rotation and demand variability. The task is performed ***product segmentation*** based on demand variability (ABC Analysis).

You can use this information as a reference for ABC Analysis: <https://www.cin7.com/industry-terms/abc-analysis/>

The references that are driving most of the sales.

* Class A: Very Fast Movers: top 5%
* Class B: The following 15% of fast movers
* Class C: The remaining 80% of very slow movers

**Task**

1. Use first year data of Household category to create ABC Analysis and interpret the graph.
2. How stable is the customers’ demand? (Coefficient of Variation)

To understand which products will bring planning and distribution challenges, compute the coefficient of variation of the yearly distribution of sales of each reference.

1. Discuss a few initiatives and recommendation for improving the retail business for dataset\_02.

***Hints***: (1) Provide more discount for location which has low median income based on specific event date, for example, Christmas. (2) Potential interpretation based on ABC Analysis and Demand variability.

# 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 among team members.

4. Follow the best practices for data visualization.

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 CA, TX, WI - 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

■