

## Assignment 2

# ML as a Service

---

Ronik Karki

Student ID: 24886412

10/10/2023

36120 - Advanced Machine Learning Application  
Master of Data Science and Innovation  
University of Technology of Sydney



## Table of Contents

<b>1. Executive Summary</b>	<b>2</b>
<b>2. Business Understanding</b>	<b>3</b>
a. Business Use Cases	3
<b>3. Data Understanding</b>	<b>4</b>
<b>4. Data Preparation</b>	<b>6</b>
<b>5. Modeling</b>	<b>8</b>
a. Approach 1	8
b. Approach 2	9
<b>6. Evaluation</b>	<b>10</b>
a. Evaluation Metrics	10
b. Results and Analysis for Approach 1	10
c. Results and Analysis for Approach 2	13
d. Business Impact and Benefits	14
e. Data Privacy and Ethical Concerns	14
<b>7. Deployment</b>	<b>15</b>
<b>8. Conclusion</b>	<b>16</b>
<b>9. References</b>	<b>17</b>



## 1. Executive Summary

This project is for creating a machine learning model and forecasting model to predict the revenue for the American retailer that has 10 stores across 3 different states: California (CA), Texas (TX), and Wisconsin (WI). Two different types of problems were identified and solved using the models for this project. Two different objectives and the outcomes of these models were:

1. Predicting Revenue for each item according to store for a specific date.

The main goal of this project is to predict the revenue of the specific item and store. This project helps the business to predict the prices that the items of a particular store will be making in the future. This helps the businesses to understand how much profit can be generated and if the price doesn't come as expected, then it can be analyzed by a specific store to look at the issues. So using a machine learning model like Decision Tree, the model could predict revenue for each item according to store for a specific date with decent ability. The MAE score indicating that the model's predictions are closer to the actual values on average was reduced to 2.41, and the RMSE indicating that the model's predictions are closer to the actual values and are sensitive to larger errors was 5.70. R2 score of 0.61 shows that it does have some ability to explain variance.

2. Predicting total revenue across all the stores and items for a particular date.

The main goal of this project is to predict the total revenue. By observing the trend, seasonality, and other aspects, the forecasting time series model created for this model would predict the revenue in future dates which would help the retailer to look at how much revenue the business will be making in the future and can work on improving areas if it doesn't meet the expected revenue as the retailer wants. So using a forecasting model like the Prophet model, the model had a much higher ability to predict the total revenue for the future dates with a Mean Absolute error of \$5,000 and RMSE of \$8000. This means that the model predicts quite near to the actual values as most values for the sales are in the range of above \$75,000 per day making an \$8,000 error to be considerable for good prediction. Furthermore, an R2 score of 0.797 shows that the model has been fit properly and explains the data variance by almost 80% efficiency.





## 2. Business Understanding

### a. Business Use Cases

This project can be used for two purposes. The first model can be used for predicting each item's price according to the stores and for a particular date. This model is super specific to the items, therefore the model can be used to get the expected scores of revenue that will be generated in the future. With this value, the retailer can assess the performance of the products being sold. They can either demand more products or reduce the product size if the product is less expected to be sold. This makes business decisions easier to gain more profit and fewer investments in products that won't give any benefit in the future. The second model can be used for planning the total revenue of the company. If the revenue is forecasted to be less than expected then additional business decisions on improving areas can be made which would again lead to more profit for the company.

### b. Key Objectives

As explained in section 2. a, the stakeholders for this project would be the retailer to improve the performance of the company. The main objective of this project would be to properly predict the revenue for each item as well as for the total revenue for future dates. Machine learning models and time series forecasting models trained on big data would have the ability to precisely predict the outcome of revenue on future dates and also would provide the estimated error of the model which can also be considered by the business while making the decisions.



### 3. Data Understanding

The dataset used for creating a machine learning model and forecasting model was collected from assignment 2 of the Advance MLA course. Different CSV files stored different information in the data. Initially, the data was in a wide format of 30,490 rows and 1547 columns. It was converted to a long format which had around 47 million data after the final table was created which is further explained in section 4 of this report.

The dataset had some limitations. The information regarding the column was not provided neither was the column name completely clear. The dataset lacked information about the customer segments and economic indicators which would have provided more insights to get the best out of the machine learning model.

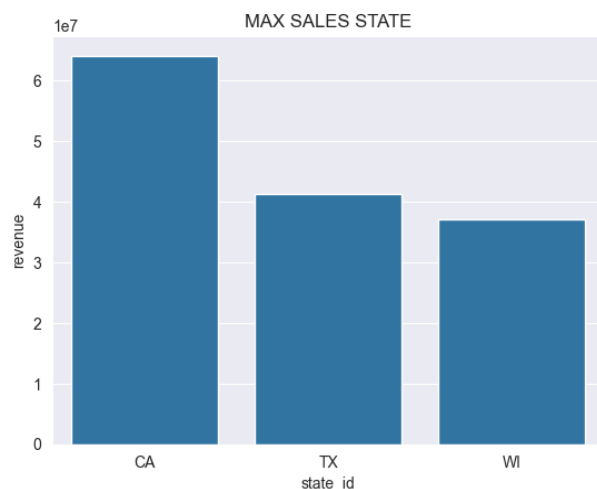


Fig 1: Revenue according to state

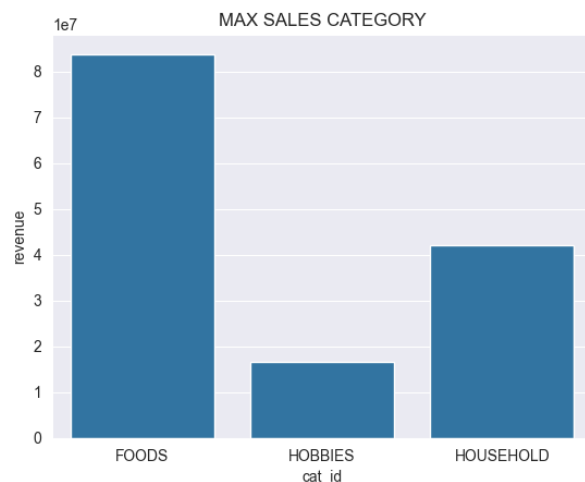


Fig 2: Revenue according to category

Dataset had the information about the 10 different stores across 3 states in the US. They are California, Texas, and Wisconsin. Among them, California had the highest number of revenue generated with Wisconsin being the lowest. Each of these states and stores had different categories sold. The categories are Foods, hobbies, and household goods. In comparison, Foods generated the most revenue followed by the household.

For the forecasting model, the data were grouped according to dates and each date had its per day total revenue across all the stores and items.

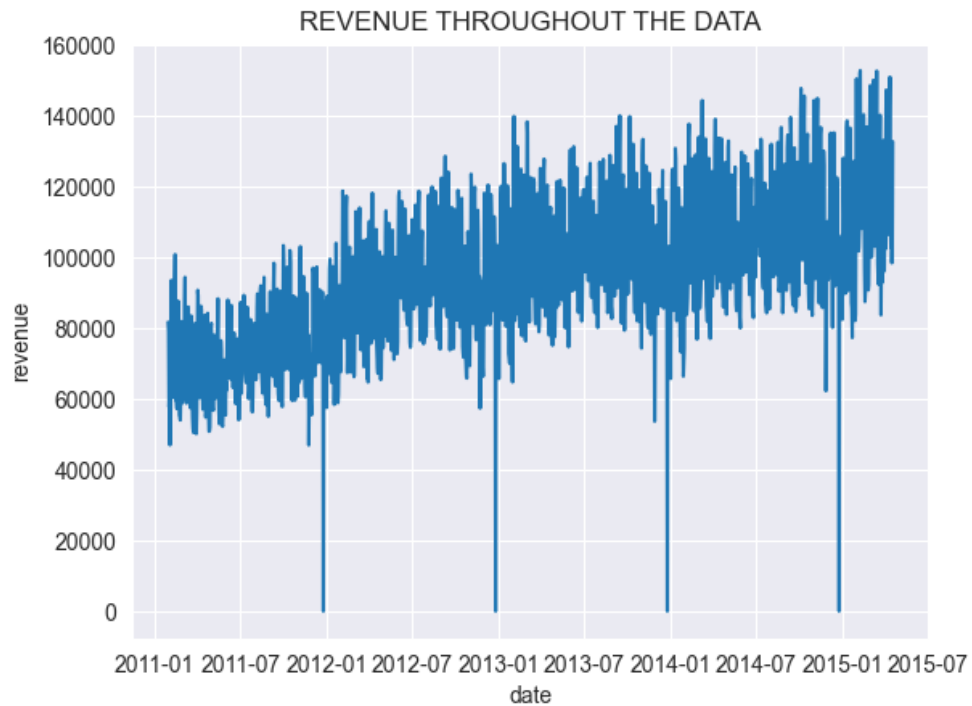


Fig 3: Revenue distribution per day

From Fig 3, it can be observed that there is a significant drop in the revenue for 1 day in each year interval. It was found out to be Christmas Day and it had an impact on the total revenue per day. Therefore capturing holiday information and fitting it to the machine learning model was one of the significant features of this dataset.

## 4. Data Preparation

For the data preparation part, there were 4 different CSV files (sales\_train.csv, calendar.csv, calendar\_events.csv, and items\_weekly\_sell\_prices.csv) which stored important information about the data set. All four CSV files were merged to create one pandas data frame for prediction. The following steps were performed for the dataset preparations:


1. Initially, the sales\_train file had information about each item ID, store ID, category ID, department ID, and state ID concerning the quantities that were sold on a specific date from the year 2011 to 2015.
2. The sales\_train file was in wide format which was melted using pandas df\_melt function to convert to the long format.
3. This was further merged with the calendar to get the date information about the items of a particular store being sold.
4. Furthermore, the merged data were again merged with the items\_weekly\_sell\_price file to get the information about the price for that particular product. In addition, each unit was multiplied by its sell price to get the revenue of each product.
5. Finally, columns like id, d, department id, wm\_yr\_wk, sales, and sell\_price were dropped as they had information captured by other columns and these columns were redundant which would just increase the complexity of the large dataset. Then this table was used for feature engineering and further model processing.

This table was used for both the models for the prediction. However, for the forecasting model, all of the columns were dropped except for the date and revenue. They were further grouped according to the dates and all the revenues were added to get the final table for the training purposes.

For the first model (i.e. predictive) three types of feature engineering techniques were used:

1. Data imputation:

There were Nan in the sell\_price column which led to the revenue having Nan values, it might be because at that period the product wasn't sold and had 0 sales. Therefore, these Nan values were imputed with 0 values which means no sales were observed on that particular date.



## 2. Ordinal encoding:

All of the categorical features were converted to integers by applying Ordinal encoding techniques and this encoder was saved for applying in the future unseen data as well.

## 3. Feature transformation:

The Date feature was converted to get specific information for the month and day which were also encoded using the ordinal encoder. The date was also converted to numerical using the DateTime to ordinal function to put the information of the date as required for prediction.





## 5. Modeling

Two different approaches were used for predicting revenue for this project. The first approach was using different machine learning algorithms to predict revenue for each item and store for a specific date. For the second approach, a forecasting model was used to predict revenue for each day across all stores and items for a specific date till the next 7 days.

### a. Approach 1

Four different machine-learning regression models were created for the first approach and the results are displayed in table 1 above. The reasons why these models were selected are explained below with their technical performances:

#### a. Linear Regression model:

Although the problem was in a hierarchical structure, a linear regression model was used to figure out if there were any linear relationships between the features and the target variable.

#### b. XGB Regressor Model:


Secondly, the XGB Regressor model was used for predicting revenue in an attempt to improve the score model using boosting techniques. The model was used without any parameters (i.e. Default settings).

#### c. Decision Tree Regressor Model:

As the problem for predicting revenue according to item and store information is a hierarchical problem, the tree model is expected to perform the best. In expectation of improving the performance and reducing the error, a Decision tree regressor was used. Initially, this model was used without any hyperparameters (i.e. Default parameters)

#### d. Tuned Decision TreeRegressor Model:

To reduce the overfitting, the Decision tree model was used and its hyperparameters were tuned. The following hyperparameters were obtained after performing the cross-validation 5 times using GridSearchCV to find out the optimal combination of the hyperparameters.

- 
- i. `max_depth = 60`
  - ii. `min_samples_split=39`
  - iii. `min_samples_leaf=18`
  - iv. `max_features=None`

## b. Approach 2

For the second approach, i.e. for forecasting models, the Prophet model was used. Due to memory limitations, information on significant event types and names of holidays was not merged from the dataset. As explained in section 3 of this report, Adding information like Holiday events (“Seasonality, Holiday Effects, and Regressors,” 2023) was significantly important for forecasting the revenue for future dates. It was difficult to add this information to other forecasting models like SARIMA, ARIMA, etc forecasting models. Therefore, the Prophet model was used for prediction as the prophet model has a built-in function to add the list of holidays according to the country names.



## 6. Evaluation

### a. Evaluation Metrics

As the problem with both of the prediction models is to predict revenue which is a continuous variable, the evaluation metrics (Brownlee, 2021) selected for assessing the performance of both models is Root Mean Squared Error (RMSE) and Mean Absolute Error(MAE). The lower value of RMSE and MAE indicates a better performance of the model. However, the score doesn't have any limit for its range. RMSE and MAE scores give different meanings according to the problem and the range of values for that particular problem. For example, if the average values of the revenue are above \$100,000 then \$5,000 RMSE and MAE errors are considerable whereas the same RMSE and MAE are much higher and unconsiderable if the average revenue is just \$6000. Furthermore, to analyze how properly the model explains the variance is measured by R2 which is also used for figuring out how well the model fits rather than assessing the performance.

### b. Results and Analysis for Approach 1

Model	Metrics	Training Scores	Testing Scores
Naive Model	MAE	-	4.38058
	RMSE	-	9.17113
	R2	-	0.0
Linear Regression Model	MAE	4.98589	4.92254
	RMSE	9.25654	9.23586
	R2	0.008	0.007
XGBRegressor Model	MAE	3.94789	4.21470
	RMSE	8.63260	9.34613
	R2	0.07238	0.05405
Decision Tree Regressor	MAE	0.0	2.91908
	RMSE	0.0	7.50662
	R2	1.0	0.33005
Decision Tree Regressor Tuned	MAE	2.16505	2.41497
	RMSE	5.16482	5.70069
	R2	0.68279	0.61362

Table 1: Scores for all regression models

#### a. Linear Regression

This model wasn't able to learn anything about the data and had very poor prediction results as shown in Table 1. In comparison to the naive models, Linear regression performed even poorly where the naive model is just the mean value predicted for all the dates. With  $R^2$  scores being close to 0, it can be concluded that the model had no idea about the data and couldn't fit the data at all.

#### b. XGB Regressor

The XGB Regressor model also couldn't learn properly from the data like the Linear Regression model. From Fig 4 below, it can be seen that the model wasn't able to predict the actual values above 100 at all. It proves that the model didn't learn from the data set.

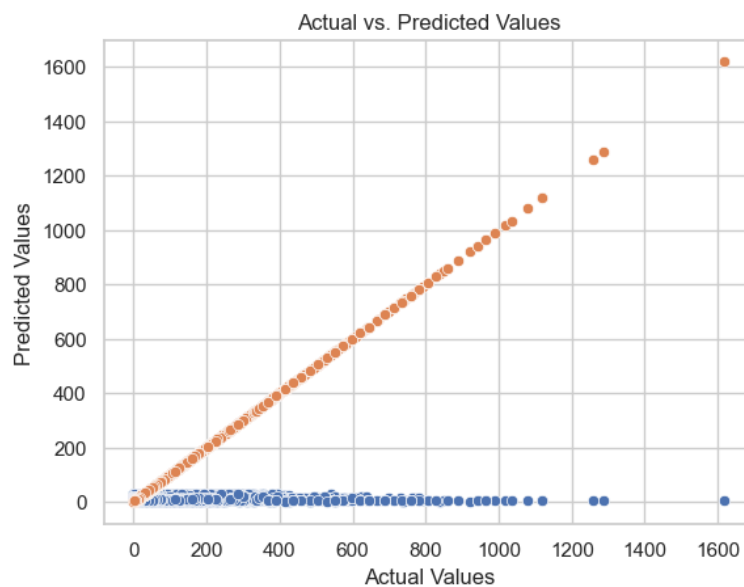


Fig 4: Actual vs predicted values for the XGB Regression model

#### c. Decision Tree

This model was more prone to the noise in the dataset and was overfitting. We can see from Table 1 that the train score has absolutely 0 errors and perfect fitting  $R^2$  scores. However, it doesn't follow the same pattern for the unseen test data. This model was trained without any parameters. Compared to the two previous models, this model somehow got the learning and therefore this model was tuned to reduce overfitting.

d. Tuned Decision Tree

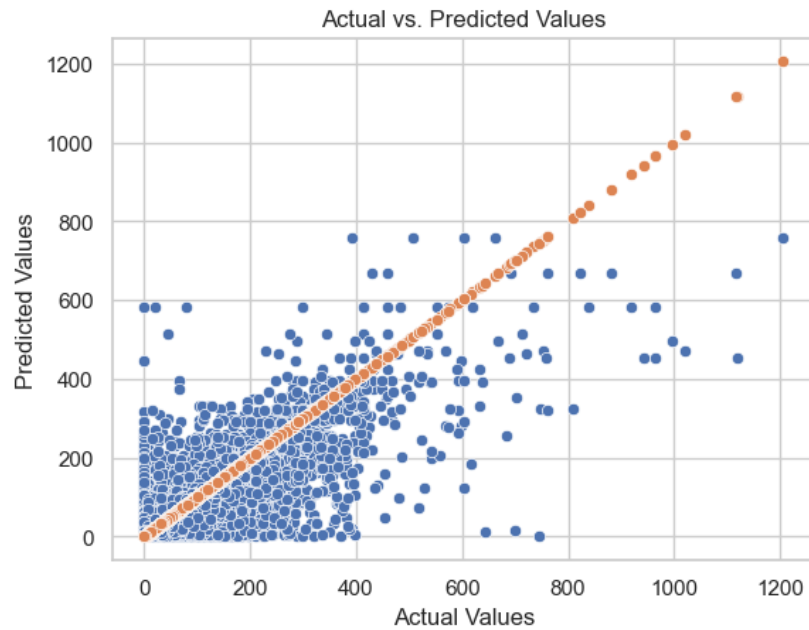


Fig 5: Actual vs predicted values for Tuned Decision Tree model.

From the above figure(i.e. Fig 5), we can see that the tuned decision tree model had good performance compared to any other model. From the Table 1 scores, we can see that the model didn't overfit and has decent performance. It has fewer errors compared to other models.

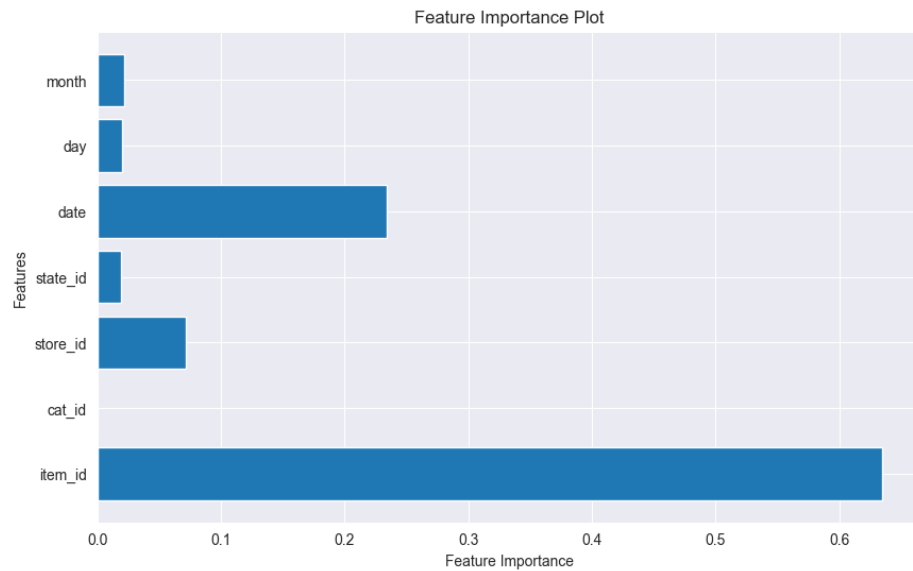


Fig 6: Feature importance for a tuned decision tree model


From the best performing model, it was found that (see Fig 6), item id had the most significant information for predicting the upcoming revenue. Date values were equally important for prediction as well.

### c. Results and Analysis for Approach 2

Model	Metrics	Testing Scores
Naive Model	MAE	15444.83224
	RMSE	18020.6735
	R2	0.0
Prophet Model	MAE	5451.2427
	RMSE	8111.2247
	R2	0.79740

Table 2: Scores for the prophet model

The prophet model was fitted with the list of holiday dates as it was important for the model to observe these dates as explained previously. For the list of holidays, the model was trained on US holidays as the data is based on US retailers. The Prophet model outperformed the naive model by a lot of improvement in the scores. The prophet model significantly reduced the MAE, RMSE, and R2 scores compared to the Naive model. It is



because the model was able to observe all the trends, seasonality, and holidays that impacted the revenue across different dates.

#### d. Business Impact and Benefits

##### a. Predictive model

The model is far better than the naive model, but the impact of prediction errors on inventory management, pricing strategies, and customer satisfaction can vary. According to each item price the MAE and RMSE scores might lead to overstocking or understocking of items which might lead to loss of business profit. This model can be used for estimations but can't be solely dependent. The regression model might not be a proper fit as it doesn't observe the trend and seasonality of the data.

##### b. Forecasting model

The model is far better than the naive model, This model can be used for estimations as it follows and shows the proper trend. The predicted value can capture significant dates as well. Therefore, the business can use this model to get the expected revenue in the future. Although the model might have errors up to \$5000 it would be considerable to figure out the total sales per day in the future.

#### e. Data Privacy and Ethical Concerns

The model has information about the revenue of the US stores. There is no personal information linked in the data. Also, the item information and the store information are converted to IDs which makes it impossible to get the real information of which store and what item price it is. Therefore it protects data privacy and removes concerns from the data.





## 7. Deployment

Using the FastApi package, the Application programming interface is created through which the two different endpoints are created. In these two different endpoints, one endpoint gives the prediction for the regression model and the other gives the prediction for the forecasting model. All of these endpoints were initially tested on the local server by creating the image file from the Docker app. Once the test was completed, the model was then deployed using the Heroku. Heroku manages Continuous Integration/ Continuous Deployment (CI/CD) pipelines to maintain the integration as well as deployment of the updated programs and models which can also be used in the future once the new data arrives and the model is updated.

The project can be accessed at: <https://protected-lake-95023-3e1d8126370a.herokuapp.com/>







## 8. Conclusion

For this project, if high-powered computation/ resources were available then, more features could be added like event type and names, and also Auto Regressive model combined with tree models might have given more trends and seasonal events information to a model which might have increased the model performances. Furthermore, models like Random Forest Regressor could be trained which might have provided more accurate results than just a Decision Tree model. However, it requires higher computational power. In the context of forecasting models, the model can be improved if more exogenous features like economic indicators, demographic data about the customers, etc were present as they would give more information which would lead the model to get more accurate predictions.





## 9. References

1. Seasonality, Holiday Effects, And Regressors. (2023, October 7). Prophet. [https://facebook.github.io/prophet/docs/seasonality\\_holiday\\_effects\\_and\\_regressors.html](https://facebook.github.io/prophet/docs/seasonality_holiday_effects_and_regressors.html)
2. Brownlee, J. (2021). Regression metrics for machine learning. MachineLearningMastery.com. <https://machinelearningmastery.com/regression-metrics-for-machine-learning/>

