Supervised ML Regression - Retail Sales Prediction

ABSTRACT

In the following project, we have applied machine learning to a real world problem of predicting retail stores sales. Such predictions helps store managers in creating effective staff schedules that increase productivity. We used popular open source programming language Python and used its libraries like NumPy, scikit-learn, pandas, matplotlib for modelling, analysis and prediction and visualization.

We have used different techniques like regression, decision tree and XGB regression. In view of nature of our problem, R squared, Root Mean Square Error (RMSE) and Mean Absolute Error is used to measure the prediction accuracy.

Keywords: NumPy, scikit-learn, machine-learning, RMSE, XGB regression, Ensemble

1. PROBLEM STATEMENT

Data provided by Rossmann gives various information of about 3,000 drug stores in 7 European countries. Currently, Rossmann store managers were tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

We are provided with historical sales data for 1,115 Rossmann stores. Our task is to forecast the "Sales" column for the test set with the help of given datasets.

2. INTRODUCTION

A major challenge for large retailers is to address the needs of the consumers more effectively on a local level, while maintaining the efficiencies of central distribution. As the demand for mass customization by consumers grows, methods focused on store level optimization increase in value.

Prediction of sales is an important application of machine learning in the retail space. Given accurate predictions, retailers can manage dynamic pricing, staff rostering and inventory so as to maximize profit and improve the customer experience.

An accurate forecast of sales allows retail outlets to answer questions such as:

- Can we use dynamic pricing to maximize our profit?
- Do we have enough stock to satisfy demand without being overstocked?
- What are the most important factors that affect sales, and how can we optimize them?

3. DATA DESCRIPTION

The dataset contains information about the stores and its sales. Two datasets are provided.

- stores_data.csv - historical data including Sales
- store.csv - supplemental information about the stores

Data fields

- Id - an Id that represents a (Store, Date) duple within the test set
- Store - a unique Id for each store
- Sales - the turnover for any given day (this is what we are predicting)

- Customers - the number of customers on a given day
- Open - an indicator for whether the store was open: 0 = closed, 1 = open
- StateHoliday - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools
- StoreType - differentiates between 4 different store models: a, b, c, d
- Assortment - describes an assortment level: a = basic, b = extra, c = extended
- CompetitionDistance - distance in meters to the nearest competitor store
- CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
- Promo - indicates whether a store is running a promo on that day
- Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2
- PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store.

4.1 LITERATURE REVIEW

Forecasting is projecting, predicting or estimating some future condition or event that is beyond an organization’s power and gives a basis for efficient planning. Forecasting is necessary for several situations of the modern business and its proper working. Organizations must make plans that will be efficient at some point in the future. And to do this they require information and data about current circumstances. It is very unfortunate that though forecasting is an important aspect yet its progress in many field or research and development has been
limited.

In the past decade Machine Learning have emerged as a technology with a great promise for identifying and modeling data patterns that are not easily described by traditional statistical methods in a field as diverse as cognitive science, computer science, electrical engineering and finance.

Example- studies in the "finance literature evidencing predictability of stock returns by means of linear regression can be improved by a neural network. Machine Learning have also been increasingly used in management, marketing and retailing. The types of applications include market response forecasting.

In this particular project we will give the following business insights to the owner

- What is the extend to which sales performance is influenced by factors like: promos, school and state holidays, competition distance ,competition open month. locality and seasonality.
- What model is appropriate to predict sales?

4.2 PROBLEM FORMULATION

Rossmann store managers had to predict the daily sales and the number of customers for up to six weeks in advance. What is the extent to which sales performance is influenced by factors like: promos, school and state holidays, competition distance, competition open month, locality and seasonality?
As there are so many of individual who try to forecast sales based on their unique sets of circumstances, the accuracy of such forecasts was rather varied. So, our task was to make an efficient machine learning model that would predict the sales for 1,115 stores across Germany using which store managers would be able to create effective staff schedules to increase their productivity and sales turnover.

4.3 FUTURE SCOPE OF THE PROJECT

Our model will help local retailers to spike their business in the following ways: -

1. It will them to decide marketing strategies.
2. It will help them preparing the budget and for setting financial policies.
3. With effective sales forecast it is feasible to obtain an average estimate of everything in such a way that the average manpower and plant capacity is fully utilized during the entire time period. Thus, the forecasting enables to overcome seasonal variations.
4. It helps in stocks organizing and prevents the risk of both the over-stocking and under stocking.
5. with the help of forecasts, we can find out which product provides more profit and which product's manufactured should be stopped.

We believe, every business will at some point in the future consider forecasting their sales for the upcoming challenges. The role of Al and Deep Learning will not just be limited to technical use but will be used in every sphere of life.

5.1 TOOLS AND TECHNOLOGIES USED

- Languages to be used: Python
- Study Focus on: Data Analysis, Machine Learning
- Tools to be used: Google collab

5.2 STEPS INVOLVED

- Loading our dataset and importing all the useful libraries
- Dealing with null values and filtering the data
- Adding additional data field to make proper analysis
- Exploratory Data Analysis
- Merging of Datasets
- Encoding of categorical columns
- Feature Selection
- Standardization of features
- Machine Learning Data Modeling (for our Prediction)

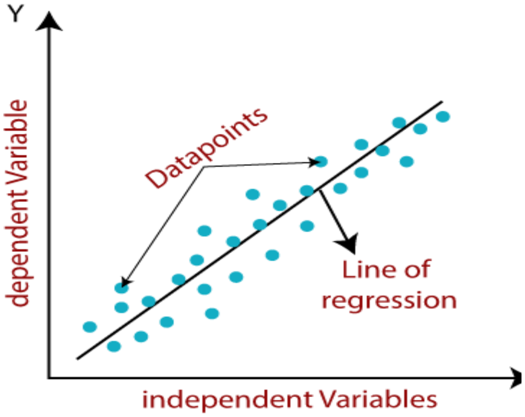
6. ALGORITHMS

1. **Linear Regression:**
   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task**.** Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship

between dependent and independent variables they are considering, and the number of independent variables getting used.

In a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

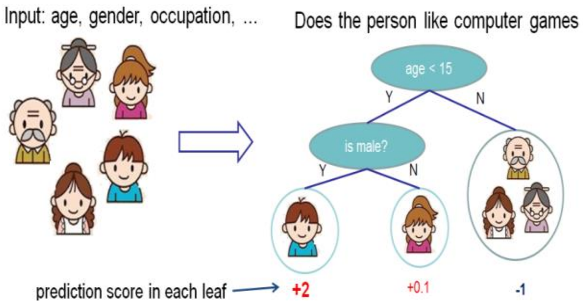


2. **XG Boost-**

To understand XG Boost we have to know gradient boosting beforehand.

Gradient Boosting-

Gradient boosted trees consider the special case where the simple model is a decision tree


Input: age, gender, occupation, ...
Does the person like computer games
age < 15
Y
N
is male?
Y
N
prediction score in each leaf
+2
+0.1
-1

In this case, there are going to be 2 kinds of parameters P: the weights at each leaf, w, and the number of leaves T in each tree (so that in the above example, T=3 and w=[2, 0.1, -1]).

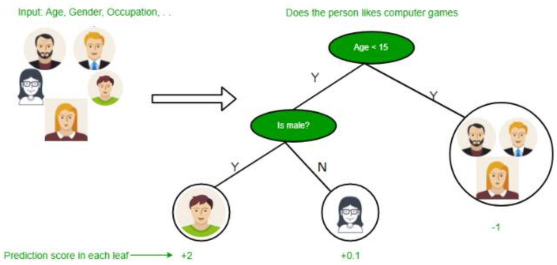
When building a decision tree, a challenge is to decide how to split a current leaf. For instance, in the above image, how could I add another layer to the (age > 15) leaf? A 'greedy' way to do this is to consider every possible split on the remaining features (so, gender and occupation), and calculate the new loss for each split; you could then pick the tree which most reduces your loss.

XG Boost is one of the fastest implementations of gradient boosting. trees. It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XG Boost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

3. **Decision Tree**

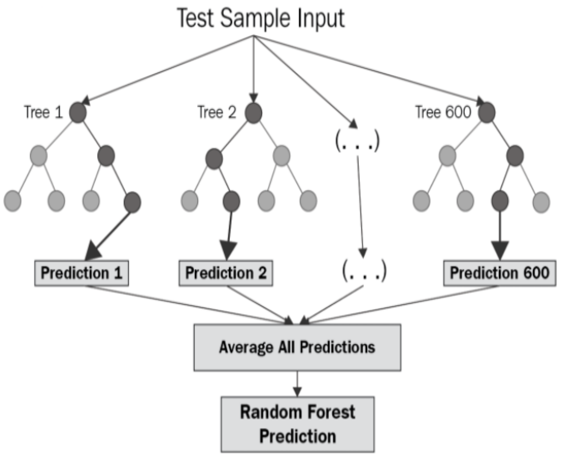
Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

We can represent any boolean function on discrete attributes using the decision tree.


Input: Age, Gender, Occupation, . . .
Does the person likes computer games
Age < 15
Y
Y
Is male?
Y
N
-1
Prediction score in each leaf
+2
+0.1

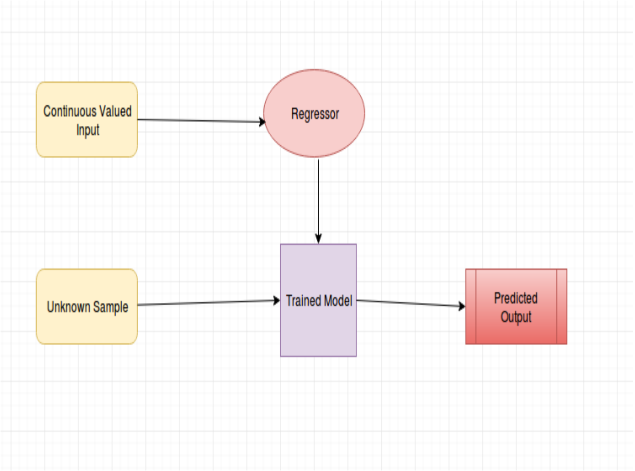
4. Random Forest:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

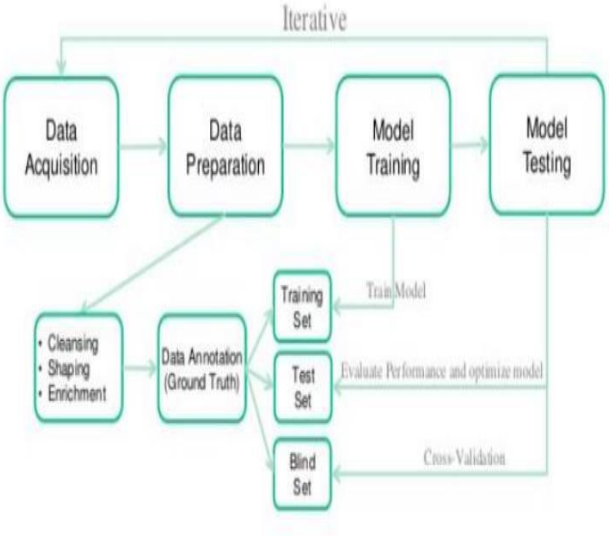


7. IMPLEMENTATION

The goal of Regression is to explore the relation between the input feature with that of the target value and give us a continuous valued output for the given unknown data.



- **Working Flowchart**



- **Testing**

Index No	Model	R^2	Mean Absolute Error	Root Mean Square Error
1	Linear Regression	0.901151	871.658135	1213.563087
2	XG Boost Regression	0.858944	916.319565	1449.681324
3	Decision Tree	0.972031	385.566520	645.524410
4	Random Forest	0.986423	271.104453	449.751225

It is clear from our table that Random Forest Model gives us the least error.

8. CONCLUSION

Out of the four methods, Random Forest proved to be the most accurate, achieving a R2_Score of 0.986449, MAE of 270.752379 and RMSE of 449.331705. While it has the lowest error of all methods, it requires more work than the other three approaches and hence, consumes more time to produce results.

So, now we can say that the Rossmann store person can now implement the Random Forest Model and utilize the feature importance data for predicting the sales for next six months