**Rossmann Sales Predication**

Pravin. M, Data Science Trainee, Almabetter, Bangalore

**Abstract:**

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are 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.

Rossmann's challenge to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable store managers to create effective planning on supply chain and logistics and also they can make staff schedules that increase productivity and motivation.

By helping Rossmann create a robust prediction model, we will help store managers stay focused on what’s most important to them: their customers and their teams.

**Keywords:**

* Rossmann Sales Prediction,
* Sales Forecast,
* Future Sales Prediction,
* Machine Learning,
* Supervised Learning,
* Hyper parameter Tuning,
* Randomized Search CV,
* Linear Regression,
* Lasso Regression,
* Ridge Regression,
* Decision Tree Regressor and
* Random Forest Regressor.

**Problem Statement**

Data provided with historical data for 1,115 Rossmann stores. The task is to forecast the “Sales” column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.

**Business Goal**

Rossmann's challenge to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable rossmann to manage their supply chain effectively and make sure that every store are supplied with required product as per prediction. For retail chain to run seamlessly its Logistic network needs to be function effectively. Forecast also enable store managers to create effective planning on supply chain and logistics and also they can make staff schedules that increase productivity and motivation.

**Introduction**

Dirk Rossmann GmbH is one of the largest drug store chains in Europe with around 56,200 employees and more than 4000 stores across Europe. In 2019 Rossmann had more than €10 billion turnover in Germany, Poland, Hungary, the Czech Republic, Turkey, Albania, Kosovo and Spain.

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are 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.

**Dataset Description**

Rossmann Stores Data.csv - historical data including Sales

store.csv - supplemental information about the stores

* 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 you 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.

**Data Wrangling**

* **Missing Values Imputation**

Rossmann Stores Data.csv – has no missing values. No future imputation is required.

Store.csv – has missing values in the features such as CompetitionDistance, CompetitionOpenSinceMonth, CompetitionOpenSinceYear, Promo2SinceWeek, Promo2SinceYear, and PromoInterval. The Imputation of these variables are involves analysing the nature of each variables. As every variables have distinct values. So imputing these variables by statistical means will create some sort of bias while modelling. So to avoid such bias nature the missing values in these variables are treated with zero.

* **Merging the Dataset**

After imputing missing values both the datasets can be merged and formed into new dataset.

* **Transforming Dependent Variable**

The Dependent variable for this prediction is Sales. Some stores were closed temporarily due to refurbishment. So there is no sales happen in those stores. We need to remove the stores which is closed for refurbishment.

* **Transforming Independent Variable**

The Independent variable State Holiday has 4 categorical values such as 0, a, b, c. And we need to convert number 0 to string 0. And need to convert a = ph, b = eh and c = ch.

**Exploratory Data Analysis**

After data cleaning up. The dataset is ready for exploratory data analysis. Exploratory Data Analysis for the rossmann sales prediction consists of 6 six sections.

1. **Correlation Analysis**

Under correlation analysis part. We’ll look deep into how each variables are correlated to each other variable.

1. **Variable Analysis**

Variable Analysis involves looking into the nature of variables based on its distribution like whether the variable is discrete or continuous in nature. Further analysis will be based on the distribution of the variable only. Also analysis involves the counts of categorical values, and skewness for continuous values.

1. **Sales vs Customer**

Sales vs Customer involves analysing each variables and their contributions towards sales and customers of rossmann. It also particularly deals with categorical variable with multiple values and its contribution towards sales and customers.

1. **Weekly Sales Analysis**

Weekly Sales Analysis involves analysing the sales with all other variables on daily basis. This analysis is important because it will show us how each variable is contributing for sales on daily/weekly basis.

1. **Customer Base Analysis**

Customers are inevitable in retail ecosystem. So customer analysis will give the overall perspective for sales forecast. And the Customer Base Analysis involves looking at the components of the stores which brings customer back again to the stores.

1. **Sales Base Analysis**

Sales Base Analysis involves analysis component of the stores which contributing significantly for the increase in volume of sales.

**Pre-processing**

Pre-processing is involves dropping off the variables which has no significant contribution for prediction the sales. One Hot Encoding is used to encode the categorical variable into 0’s and 1’s. The Wholesome data is then chopped off into train and test set by train test splitter. As there is some sort of skewness in the sales data and customer data to avoid bias nature on prediction. The dataset needs to be standardized. After standardizing it can be used for modelling.

**Modelling**

Modelling involves five prominent process. All the five process is common for all the Machine Learning models.

1. **Selection**

Selection involves choosing the ML Algorithm based on the type of problem which we’re going to solve. Broadly there are two classification of ML Algorithms. They are Regression and Classification. Regression is used for predicting the dependent variable which of continuous type. And classification is used for predicting dependent variable of discrete type. Here we have selected five regression based algorithms. They are

Linear Regression,

Lasso Regression,

Ridge Regression,

Decision Tree Regressor and

Random Forest Regressor.

1. **Fitting**

After selection the right algorithm. The next process is fitting. Here as the dependent variable nature is of continuous nature Regression algorithms are chosen. After this training dataset consist of both independent and dependent variable is fitted to the model and allow the model to learn from training set.

1. **Predicting**

After training of model accomplished. It is now fitted with only independent variables of testing set. The model start to predict the dependent variable based on its previous gained knowledge on training set.

1. **Evaluating**

After prediction process accomplished. The Predicted results are compared with the original results. Analysing deviation of predicted values from the actual values. This analysis is also known as error analysis. Some error analysing metrics are

MSE – Mean Squared Error,

MAE – Mean Absolute Error,

RMSE – Root Mean Squared Error,

RMSLE – Root Mean Squared Log Error and

R2 Score – for variance analysis.

1. **Hyper Parameter Tuning**

After evaluation of the model. To increase the performance of the model. We need to tweak the hyper parameter involved in the model. To get best hyper parameter for the model is by performing cross validation. Cross Validation involves making the whole dataset into folds for example take 5 folds. And it will swap the train and testing set between those folds and it will fits all the hyper parameters one after the other on further iterations. It will find best parameters for the model.

**Results Comparison**

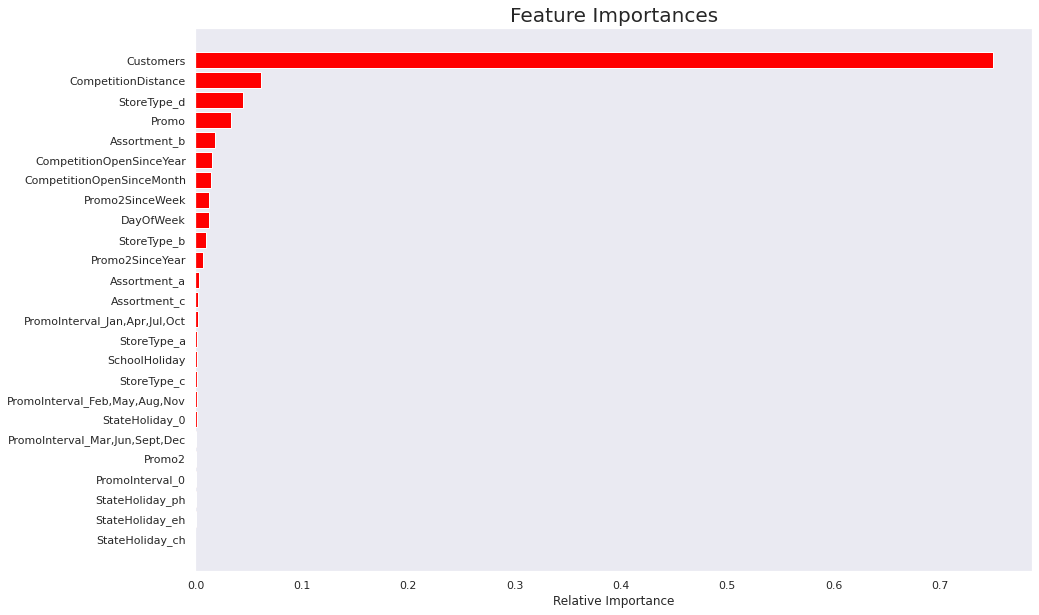
Based on the evaluation metric score and bias variance nature of the each model are also taken into account and compared with each other.



By comparing from above chart. We could see that Random Forest Regressor has performed better on testing set with low error value.

**Feature Importance**

Feature Importance involve looking deeply into the features/variable which has contributed for the prediction.



**Conclusion**

* Comparing results of all Machine Learning model based on the evaluation metrics.
* Linear Regression, Lasso Regression and Ridge Regression mostly underperform when it is used for prediction.
* As these linear models will work fine when the correlation between each variable is high and either or other the dependent variable is correlated to the each other’s.
* These Linear models are high bias and low variance in nature.
* Decision Tree Regressor performs well in training set and not so well in testing set.
* But it is working better than all the linear models.
* Decision Trees are Low bias and high variance in nature.
* Random Forest Regressor works well on both training and testing set.
* And Random Forest Regressor has low bias and high variance (compared to decision tree it is low).
* So we can use Random Forest Regressor model and it has low MSE, MAE, RMSE, and RMSLE values in testing set.