Retail Sales Prediction

# Team Members:-

# Pranav Balpande

# Kartik Dhande

# Kartik Pisudde

# Sanket Bhosale

# Puja Nehare

# Data Science Trainee

# AlmaBetter, Bangalore

**Abstract:**

Sales forecasting refers to the process ofestimating demand for or sales of a particularproduct over a specific period of time.

Businesses use sales forecasts to determine whatrevenue they will be generating in a particulartimespan to empower themselves with powerfuland strategic business plans. Important decisionssuch as budgets, hiring, incentives, goals,acquisitions and various other growth plans areaffected by the revenue the company is going tomakeinthecomingmonthsandfortheseplanstobe as effective as they are planned to be it isimportant for these forecasts to also be as good.

The sales forecasts are also different from thesales-goals a company has. Sales-goals is what acompany wants to happen to execute their futureplans for the business. On the other hand salesforecasts are what is going to happen on the basisof past records, data, trends and variousimprovement measures taken.

The work here predicts the sales for a drug storechain in the European market for a time period ofsix weeks and compares the results of machinelearning algorithms.

***Keywords:EDA,Correlation,DecisionTreeRRandom Forest, Regression, Forecasting***

# ProblemStatement:

Rossmann operates over 3,000 drug stores in 7

European countries. Currently, Rossmann storemanagers are tasked with predicting their dailysales for up to six weeks in advance. Store salesare influenced by many factors, includingpromotions, competition, school and stateholidays,seasonality,andlocality.Withthousandsof individual managers predicting sales based ontheir unique circumstances, the accuracy of resultscan be quite varied.

You are provided with historical sales data for1,115Rossmannstores.Thetaskistoforecastthe"Sales" column for the test set. Note that somestores in the dataset were temporarily closed forrefurbishment.

# Introduction:

The interest for a product continues to changeoccasionally. No business can work on itsmonetary growth without assessing client interestandfuturedemand ofitemsprecisely.

Sales forecasting refers to the process ofestimating demand for or sales of a particularproduct over a specific period of time.

For a good sales forecast, it is extremely importantto get a good dataset as well. Forecasts heavilydepend on the past records, trends and patternsobserved for sales of a particular store. Thevariations could be due to a number of reasons.

Talking from a business’s point of view, thesesalesforecastsaredoneconsistentlytoimprove

their sales forecasting models as they directlyimpact their decision making process, goals,plans and growth strategies.

In this Retail Sales Prediction, machine learningmodelsarecreatedthatpredictsalesofthese1115 drug stores across the European market andcompare the results of these models. In additionto this, an effort has been made to analyze andfindallthefeaturesthatarecontributingtohighersales and the features which are leading to lowersales, so that improvement plans can be workedupon.

# Approach:

The approach followed here is to first check thesanctity of the data and then understand thefeatures involved. The events followed were inour approach:

## Understanding the business problem and the datasets

* **Data cleaning and preprocessing-**finding null values and imputing themwith appropriate values.

Converting categorical values intoappropriate data types and merging thedatasets provided to get a final dataset towork upon.

* **Exploratory data analysis-** ofcategorical and continuous variablesagainstour target variable.
* **Data manipulation-** feature selectionand engineering, feature scaling, outlierdetection and treatment and encodingcategorical features.
* **Modeling**- The baseline model-Decision tree was chosen consideringour features were mostly categoricalwithfewhavingcontinuousimportance.

## Model Performance and Evaluation

* **StorewiseSalesPredictions**
* **Conclusion and Recommendations**

# UnderstandingtheData:

First step involved is understanding the data andgettinganswerstosomebasicquestionslike;Whatis the data about? How many rows or observationsare there in it? How many features are there in it?What are the data types? Are there any missingvalues? And anything that could be relevant anduseful to our investigation. Let’s just understandthe dataset first and the terms involved beforeproceedingfurther.

Our dataset consists of two csv files, the first consists of historical data with 1017209 rows orobservations and 9 columns with no null values.The second dataset was supplementaryinformation about the stores with 1115 rows and10 columns and a lot of missing values few columns. The data types were of integer, float and object in nature.

Let’sdefinethefeaturesinvolved:

* **Id-**anIdthatrepresentsa(Store,Date)duple within the set
* **Store-**a uniqueId foreachstore
* **Sales-**theturnoverforanygivenday(DependentVariable)
* **Customers-**thenumberofcustomersona given day
* **Open-**anindicatorforwhetherthestorewas open: 0 = closed, 1 = open
* **State Holiday -** indicates a state holiday.Normally all stores, with few exceptions,are closed on state holidays. Note that allschools are closed on public holidays andweekends. a = public holiday, b = Easterholiday,c=Christmas,0= None
* **School Holiday -** indicates if the (Store,Date) was affected by the closure ofpublic schools
* **StoreType-**differentiatesbetween4differentstore models:a, b, c,d
* **Assortment** - describes an assortmentlevel: a = basic, b = extra, c = extended.An assortment strategy in retailinginvolves the number and type ofproducts that stores display for purchaseby consumers.
* **Competition Distance**- distance in meters to the nearest competitor store
* **Competition Open Since[Month/Year]** - gives the approximate year and month ofthe time the nearest competitor wasopened
* **Promo**-indicateswhetherastoreisrunning a promo on that day
* **Promo2** - Promo2 is a continuing andconsecutivepromotionforsomestores:0

= store is not participating, 1 = store isparticipating

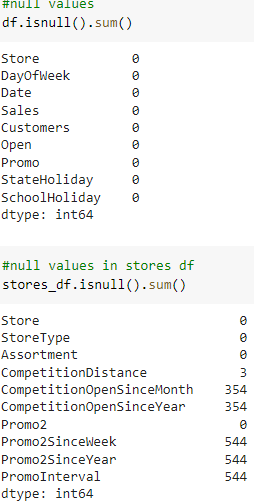
* **Promo2Since[Year/Week]-**describesthe year and calendar week when thestore started participating in Promo2
* **Promo Interval**- describes the consecutive intervals Promo2 is started, naming the months the promotion isstarted anew. E.g. "Feb, May, Aug, Nov" means each round starts in February,May, August, November of any givenyear for that store.

**Data Cleaning and Preprocessing:**Handling missing values is an important skill inthe data analysis process. If there are very fewmissing values compared to the size of thedataset, we may choose to drop rows that have

Missing values. Otherwise, it is better to replacethem with appropriate values.

It is necessary to check and handle these valuesbefore feeding it to the models, so as to obtaingood insights on what the data is trying to say and make great characterisation and predictions which will in turn help improve the business'sgrowth.

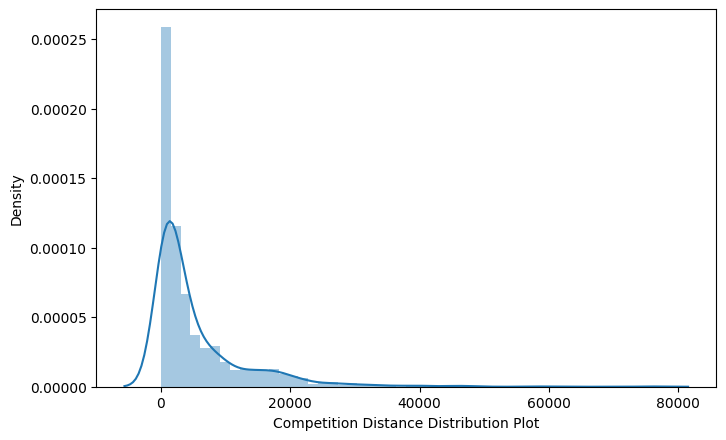
The historical records dataset had no null values.



The dataset had a lot of nulls in the followingcolumns:

* CompetitionOpenSinceMonth
* CompetitionOpenSinceYear
* Promo2SinceWeek
* Promo2SinceYear
* PromoInterval
* ‘CompetitionDistance’ - CompetitionDistance is the distance in meters to thenearest competitor store.

The Competition Distance distributionplot shows the distances at whichgenerally the stores are opened



It seems like most of the values of theCompetitionDistance are towards the leftand the distribution is skewed on theright. Median is more robust to outliereffect hence median was imputed in thenull values.

Right skewed distributions occur when the longtail is on the right side of the distribution alsocalled as positive skewed distribution whichessentially suggests that there are positiveoutliers far along which influences the mean.

It seems like most of the values of theCompetitionDistance in the column are between0-10kms. Consequently, the longer tail in anasymmetrical distribution pulls the mean awayfrom the most common values. The mean isgreaterthanthemedian.Themeanoverestimates the most common values in thedistribution and hence median is used in thiscase,itismorerobusttooutliereffectandhencemedian is used to impute the missing values inthis feature.

* CompetitionOpenSinceMonth- gives the

approximate month of the time the nearestcompetitor was opened. The mode of thecolumn is used to impute the missingvalues in the column as it gives the mostoccurring month.

* CompetitionOpenSinceYear-gives theapproximate year of the time the nearestcompetitorwasopened.Themodeofthecolumn is used to impute the missingvalues in the column as it gives the mostoccurring month.
* Promo2SinceWeek,Promo2SinceYearandPromoInterval are NaN wherever Promo2is0orFalseascanbeseeninthefirstlook of the dataset. They are replaced with0.

Lastly before proceeding further, the two datasetswere merged on the common column of ‘Store’ toget everything together for the analysis.

# Exploratory DataAnalysis:

Exploratory data analysis is a crucial part of dataanalysis. It involves exploring and analyzing thedataset given to find out patterns, trends andconclusions to make better decisions related tothe data, often using statistical graphics and otherdata visualization tools to summarize the results.The visualization tools involved in theinvestigation are python libraries- matplotlib andseaborn.

The goal here is to explore the relationships ofdifferent variables with ‘Sales’ to see what factorsmight be contributing to the high and low salesnumbers.

## Approach:

There are two kinds of features in the dataset:CategoricalandNonCategoricalVariables.

Categorical- A categorical variable is a variablethat can take on one of a limited, and usually

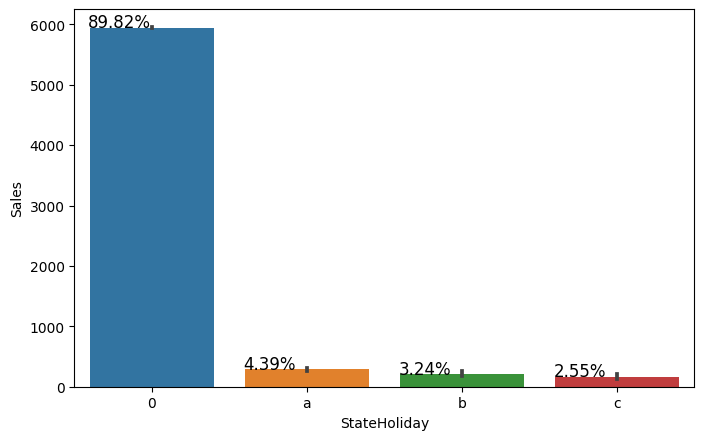
fixed, number of possible values putting aa particular category to the observation.

Non Categorical- A non categorical orcontinuous variable is a variable whose value isobtained by measuring, i.e., one which can takeon an uncountable set of values.

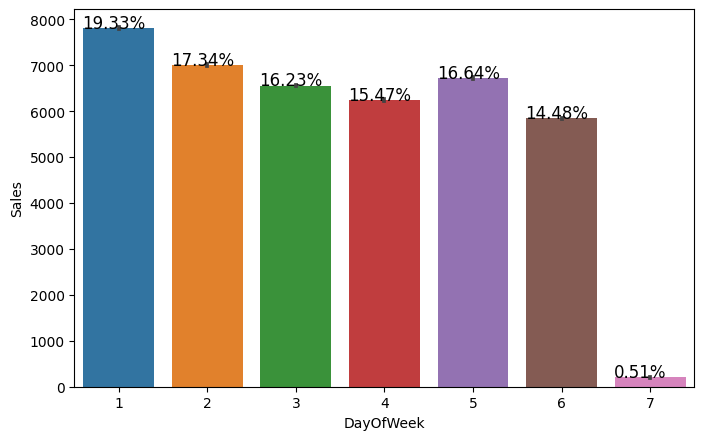
Both of them are analyzed separately.Categorical data is usually analyzed throughcount plots and barplots in accordance with thetargetvariableandthatiswhatisdoneheretoo.On the other hand Numeric or Continuousvariables were analyzed through distributionplots, box plots and scatterplots to get usefulinsights.

## Hypotheses

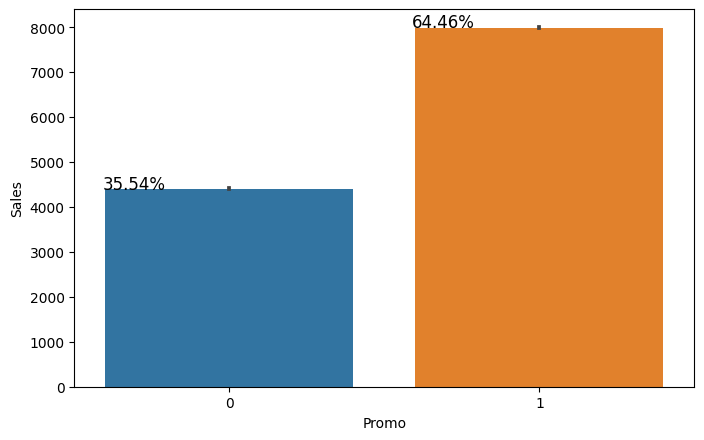
Just by observing the head of the dataset andunderstanding the features involved in it, thefollowing hypotheses could be framed:

* There's a feature called "DayOfWeek"with the values 1-7 denoting each day ofthe week. There would be a week offprobably Sunday when the stores wouldbe closed and we would get low overallsales.
* Customers would have a positivecorrelation with Sales.
* The Store type and Assortment strategyinvolved would be having a certaineffect on sales as well. Some premiumhigh quality products would fetch morerevenue.
* Promotion should be having a positivecorrelation with Sales.
* Some stores were closed due torefurbishment, those would generate 0revenue for that time period.
* Stores are influenced by seasonality,probably before holidays sales would behigh.

## Categorical Insights:

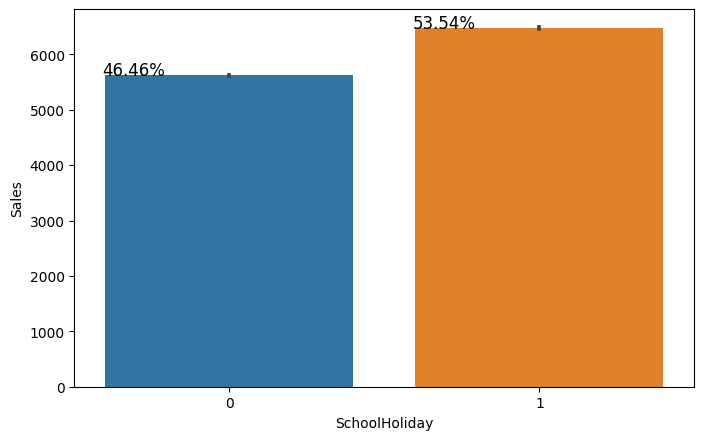


Here it can be deduced that there were more saleson Monday, probably because shops generallyremain closed on Sundays which had the lowestsales in a week. This validates the hypothesisabout this feature.

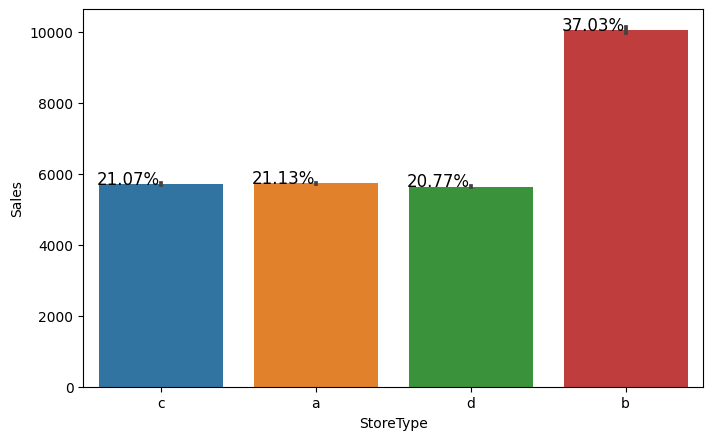


Promotion has a positive effect on SalesindicatinghighsalesforstoreswithPromo=1.

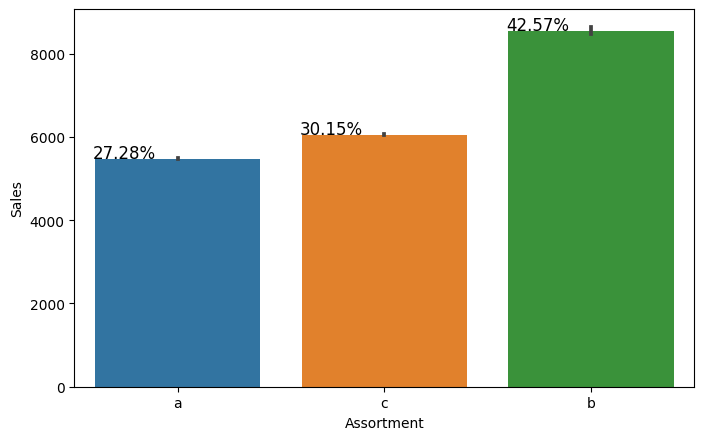
Sales were low whenever there was a StateHoliday indicating only a few stores were openon these days.



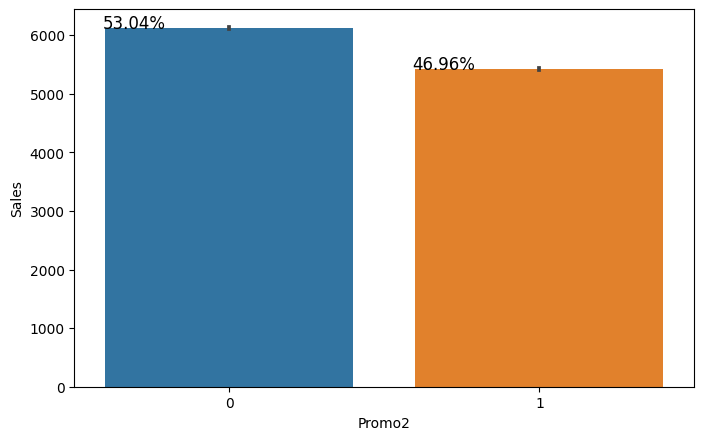
Sales were high on an average on SchoolHolidays indicating School Holidays weren’tcompulsory by the law and comparatively moresales were recorded on holidays.



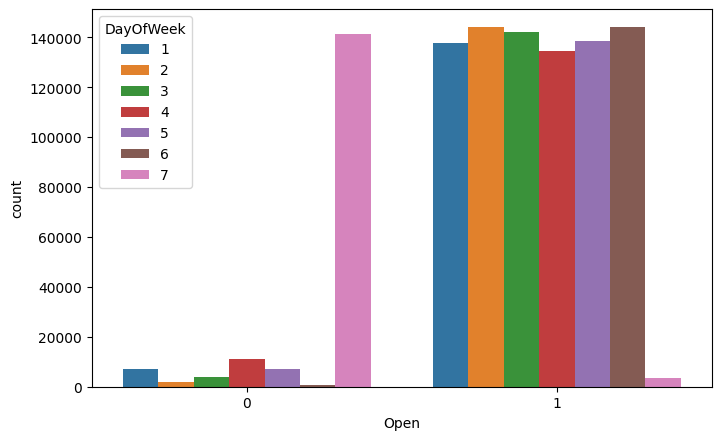
A bar plot represents an estimate of centraltendency for a numeric variable with the heightof each rectangle. The store type b has thehighest sales on an average.



Assortment type b has the highest sales on anaverage.

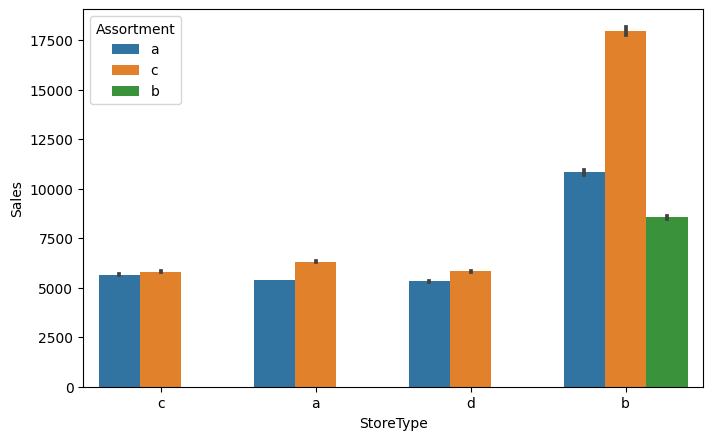


With Promo2, slightly more sales were seenwithout it which indicates there are many storesnot participating in promo.

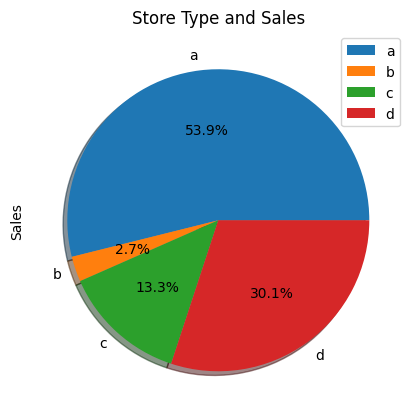


This is a count plot of open shops according tothe day of the week. It's clear that the number ofshops open on Sundays were very less and hencelow sales. Some shops were closed on weekdaysas well accounting to the stores closed due to

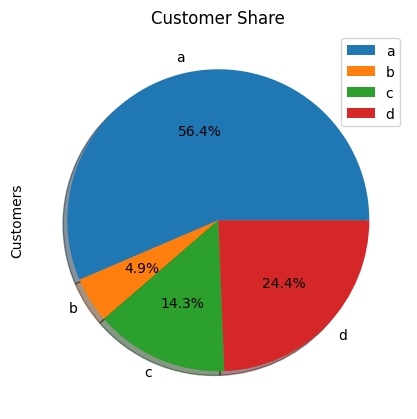
refurbishment or holidays.

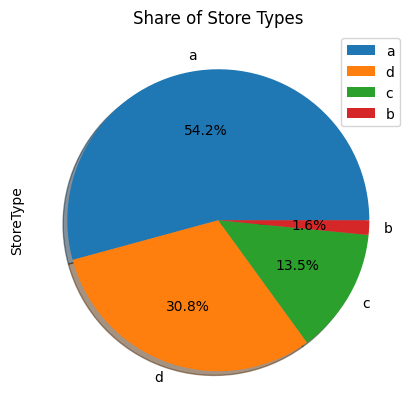


The above bar plot shows that the store types a,c and d have only assortment level a and c. Onthe other hand the store type b has all the threekinds of assortment strategies, a reason whyaverage sales were high for store type b stores.



When plotting a pie chart for the sum of sales ofthe various store types, it can be clearlyobserved that even though type a stores had themost sales, type b stores were high on anaverage.





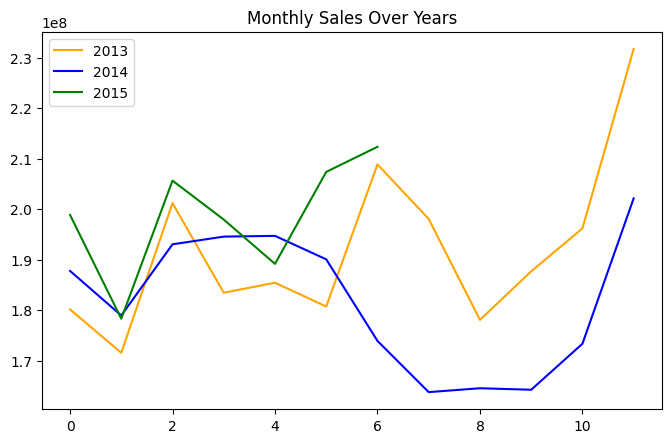
But upon further exploration it can be clearlyobserved that the highest sales belonged to thestore type a due to the high number of type astores in our dataset. Store type a and c had asimilar kind of sales and customer share.

Interesting insight to note is that store type b withhighest average sales and per store revenuegeneration looks healthy and a reason for that

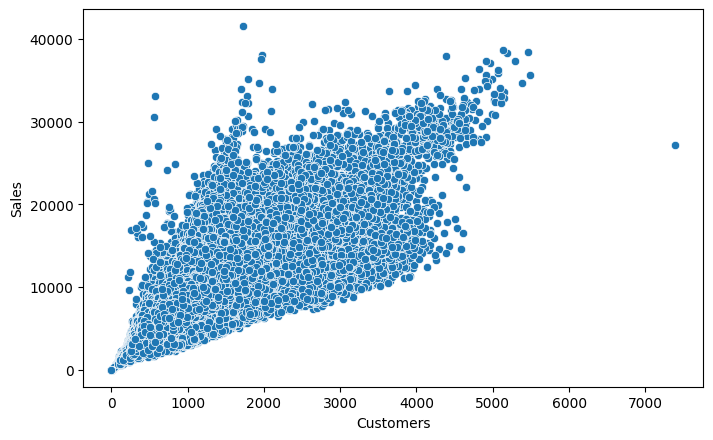
would be all three kinds of assortment strategiesinvolvedwhich wasseen earlier.

Based on the above findings it seems that thereare quite a lot of opportunities in store type 'b' &'d' as they had more number of customers perstore and more sales per customer, respectively.Store type a & c are quite similar in terms of"per customer and per store" sales numbers andjust because the majority of the stores were ofthese kinds, they had the best overall revenuenumbers. On the other hand, store type b werevery few in number and even then they hadbetter average sales than others.

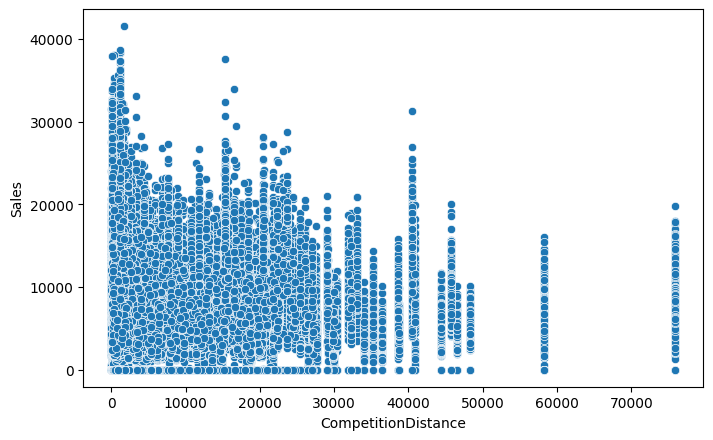
## Continuous Insights:



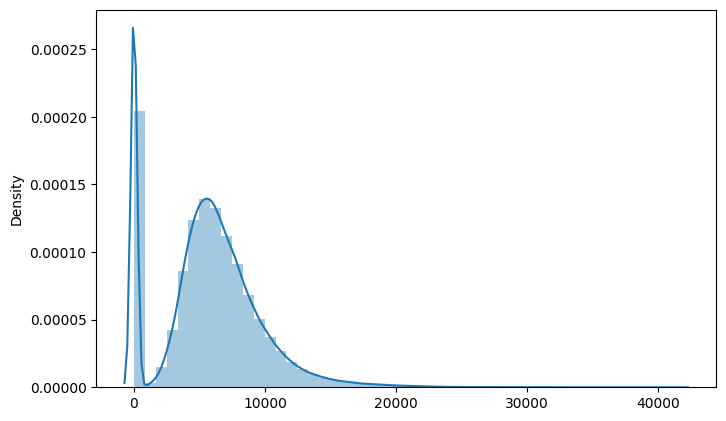
Here’s a plot of Monthly Sales over the years.Sales rise up by the end of the year before theholidays. Sales for 2014 went down there for acouple months - July to September, indicatingstores closed due to refurbishment.



Sales and Customer scatter plot showed a directpositive relation between them with a fewoutliers.



From the above scatter plot it can be observedthat mostly the competitor stores weren't that farfrom each other and the stores densely locatednear each other saw more sales. This couldindicate competition between busy locations vsremote locations.



Here’s a distribution plot of the Sales column.The drop in sales indicates the 0 sales accountingto the stores temporarily closed due torefurbishment.

## Correlation:

Correlation is a statistical term used to measurethe degree in which two variables move inrelation to each other. A perfect positivecorrelation means that the correlation coefficientis exactly 1. This implies that as one variablemoves,eitherupordown,theothermovesinthe

same direction. A perfect negative correlationmeans that two variables move in oppositedirections, while a zero correlation implies nolinear relationship at all.

Bycheckingthecorrelationthefactorsaffectingsales can be figured out.



* Day of the week has a negativecorrelation indicating low sales as theweekends, and promo, customers andopen has positive correlation.
* State Holiday has a negative correlationsuggesting that stores are mostly closedon state holidays indicating low sales.
* CompetitionDistance showing negativecorrelation suggests that as the distanceincreases sales reduce, which was alsoobservedthroughthescatterplotearlier.
* There's multicollinearity involved in thedataset as well. The features telling thesame story like Promo2, Promo2 sinceweek and year are showingmulticollinearity.

# DataManipulation:

Data manipulation involves manipulating andchanging our dataset before feeding it to variousregression machine learning models.This

involves keeping important features, outliertreatment, feature scaling and creating dummyvariablesif necessary.

## FeatureEngineering:

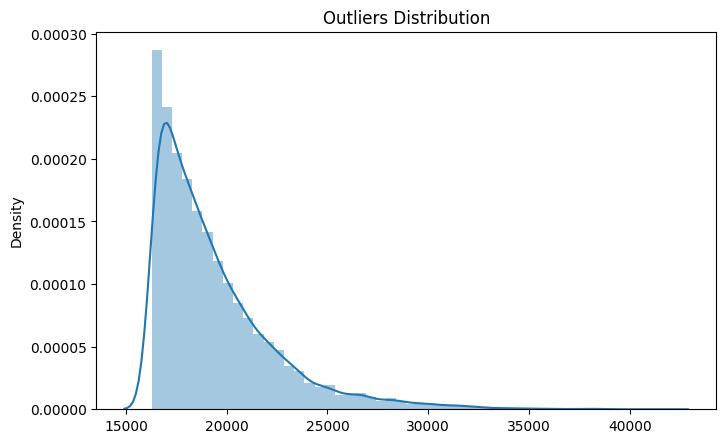
* Some stores were closed due torefurbishment and some on account ofweek off or holidays. Those stores onthosedatesgeneratedzerosalesandhenceremoving the rows was important to avoidconfusion by the algorithms and thenremoving the feature altogether because itwasn’t providing any value in predictionof the sales.
* There were features that like CompetitionOpen since Month and Year. It wascombined to count the total months sincethe nearest competition was opened.
* Promo2SinceWeek, Promo2SinceYearindicated promotion 2 opened since weekandyear.Thesefeatureswerecombinedtocount thetotal monthssince promotion2is run.
* PromoInterval indicated the months forpromotion 2 renewal. Hence, the salemonth was compared against the intervaland a new feature was created todetermine whether the promo2 wasrenewed in that month.

## OutlierDetection:

In statistics, an outlier is a data point that differssignificantlyfromotherobservations.Outlierscan occur by chance in any distribution, but theyoften indicate either measurement error or that thepopulation has a heavy-tailed distribution.

Z-score is a statistical measure that tells you howfar a data point is from the rest of the dataset. In amore technical term, Z-score tells how manystandard deviations away a given observation isfrom the mean.

z = (x-mean)/standard deviation



More than 3 standard deviations was consideredas an outlier. Exploring the outliersdataframe,some important insights were generated:

* The data points with sales value higherthan 28000 are very low and hence theycan be considered as outliers.
* The outliers had day of the week as 7 i.e.Sunday and the store type for thoseobservations were ‘b’.
* Other outliers had promotion running onthatday.
* Itcanbewellestablishedthattheoutliers are showing this behavior for thestores with promotion = 1 and store type

B. It would not be wise to treat thembecause the reasons behind this behaviorseemsfair.

* Being open 24\*7 along with all kinds ofassortments available is probably thereason why it had higher average salesthan any other store type.
* If the outliers are a valid occurrence itwould be wise not to treat them bydeleting or manipulating them especiallywhen we have established the ups anddownsofthetargetvariableinrelationtothe other features. It is well establishedthat there is seasonality involved and nolinear relationship is possible to fit. For

these kinds of dataset tree based machinelearning algorithms are used which arerobustto outlier effect.

## FeatureScaling:

Feature Scaling is a technique to standardize theindependent features present in the data in a fixedrange. It is done to prevent biased nature ofmachine learning algorithms towards featureswith greater values and scale. The two techniquesare:

Normalization: is a scaling technique in whichvalues are shifted and rescaled so that they end upranging between 0 and 1. It is also known as

Min-Max scaling. [0,1]



Standardization: is another scaling techniquewhere the values are centered around the meanwith a unit standard deviation. This means thatthe mean of the attribute becomes zero and theresultant distribution has a unit standarddeviation. [-1,1]



Normalization of the continuous variables wasdonefurther.

## One hot encoding:

For categorical variables where no such ordinalrelationship exists, the integer encoding is notenough. We have categorical data integersencoded with us, but assuming a natural order andallowing this data to the model may result in poorperformance.

Many of the features such as DayofWeek,StoreTypeand Assortments were categorical innatureandhadtobeonehotencodedtocontinue.

# Modeling:

Factorsaffectinginchoosingthe model:

Determining which algorithm to use depends onmany factors like the problem statement and thekind of output you want, type and size of thedata, the available computational time, numberof features, and observations in the data, tonamea few.

The dataset used in this analysis has:

* A multivariate time series relation withsales and hence a linear relationshipcannot be assumed in this analysis. Thiskind of dataset has patterns such as peakdays, festive seasons etc which wouldmost likely be considered as outliers insimple linear regression.
* Having X columns with 30% continuousand70%categoricalfeatures.Businessesprefer the model to be interpretable innature and decision based algorithmswork better with categorical data.

## Train-TestSplit:

In machine learning, train/test split splits thedata randomly, as there’s no dependence fromoneobservationtotheother.That’snotthecasewithtimeseriesdata.Here,it’simportanttousevalues at the rear of the dataset for testing andeverything else for training.

The latest six weeks were kept as a testing setand the rest of the historical data was used in thetraining set.

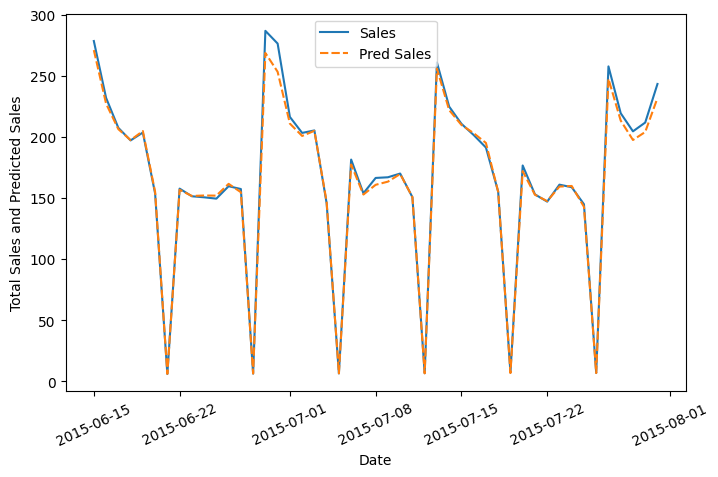
## BaselineModel-DecisionTree:

A baseline is a simple model that providesreasonable results on a task and does not requiremuch expertise and time to build. It is well

established that there is seasonality involved andno linear relationship is possible to fit. For thesekinds of datasets tree based machine learningalgorithms are used which are robust to outliereffects which can handle non-linear data setseffectively.

DecisionTreeisaSupervisedlearningtechniquethat can be used for both Classification andRegression problems. t is a tree-structuredclassifier, where internal nodes represent thefeatures of a dataset, branches represent thedecision rules and each leaf node represents theoutcome.

In a Decision tree, there are two nodes, which arethe Decision Node and Leaf Node. Decisionnodes are used to make any decision and havemultiple branches, whereas Leaf nodes are theoutput of those decisions and do not contain anyfurther branches.





The results show that a simple decision tree isperforming pretty well on the validation set but ithas completely overfitted the train set with a testR^2 of 0.91. It's better to have a much moregeneralized model for future data points.

Businesses prefer the model to be interpretable innature in order to understand the patterns andstrategize accordingly unlike any scientific

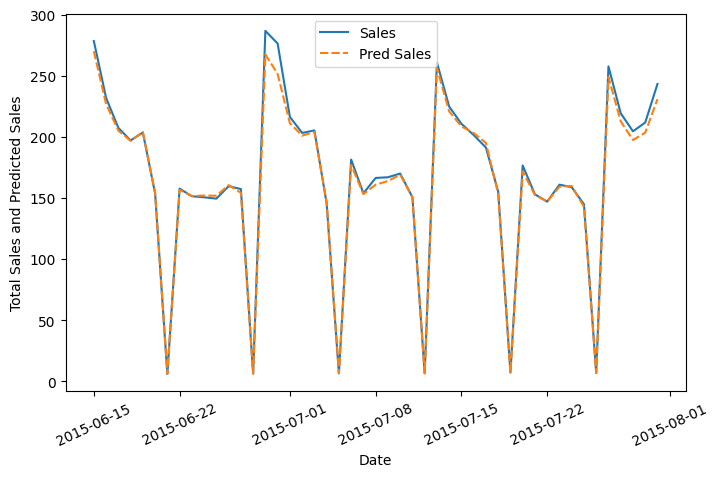
facility where the results matter much more thaninterpretability.

If interpretability is important then sticking withtree based algorithms when most of the featuresare categorical; is beneficial and using tunedHyperparameters to grow the tree deep enoughwithout overfitting.

## RandomForest:

Random forests are an ensemble learningmethod for classification and regression thatoperates by constructing a multitude of decisiontrees at training time. For regression tasks, theoutput of the random forest is the average of theresults given by most trees.

In simple terms, random forest builds multipledecisiontreesandmergesthemtogethertogetamore accurate and stable prediction.





Random Forest Regressor results were muchbetter than our baseline model with a test R^2 of0.955673.

**Random Forest Hyperparameters:**

* max\_depth- The max\_depth of a tree inRandom Forest is defined as the longestpath between the root node and the leafnode
* min\_sample\_split- a parameter that tellsthe decision tree in a random forest theminimum required number ofobservations in any given node in order tosplit it.

The default value of theminimum\_sample\_split is assigned to 2.This means that if any terminal node hasmore than two observations and is not apure node, we can split it further intosubnodes.

* max\_leaf\_nodes-Thishyperparametersets a condition on the splitting of thenodes in the tree and hence restricts thegrowth of the tree. If after splitting wehave more terminal nodes than thespecified number of terminal nodes, it willstopthesplittingandthetreewillnotgrowfurther.
* min\_samples\_leaf- This Random Foresthyperparameter specifies the minimumnumber of samples that should be presentin the leaf node after splitting a node.
* n\_estimators- the number of trees
* max\_sample (bootstrap sample)-Themax\_sampleshyperparameter determineswhat fraction of the original dataset isgiven to any individual tree.
* max\_features-Thisresemblesthenumberof maximum features provided to eachtree in a random forest.

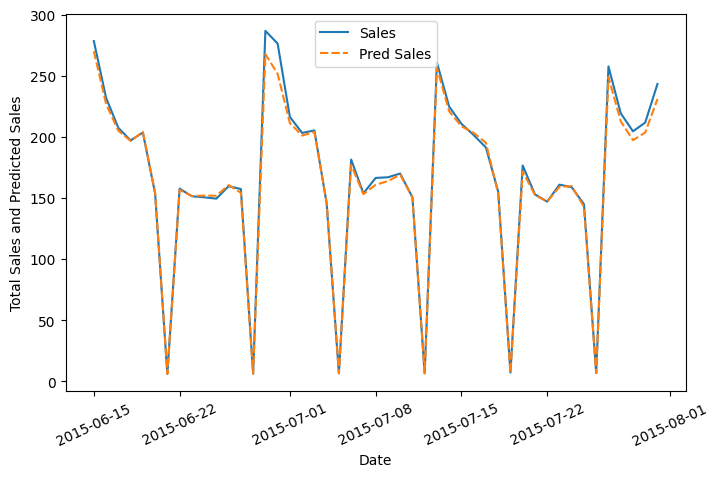
Randomized searchcv searches on hyperparameters to fit and score various models and getthebestestimator.IncontrasttoGridSearchCV,

not all parameter values are tried out, but rathera fixed number of parameter settings is sampledfrom the specified distributions. The number ofparameter settings that are tried is given byn\_iter.

## RandomForestHyperparameterTunedModel :

The maximum R^2 was seen in the tunedRandom Forest model with the value 0.955878which was only 0.021% improved from a simplerandom forest model.

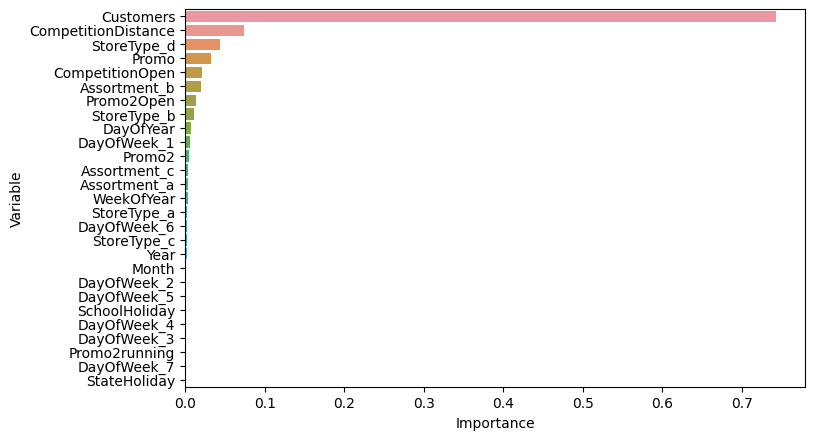
This indicates thatall the trendsand patternsthat could be captured by these models withoutoverfitting were done and the maximum level ofperformance achievable by the model wasachieved.





## RandomForestHyperparameterTunedModel

Feature Importance:



The most important features in predicting theSales were Customers, CompetitionDistance,StoreTypeD andPromo.

# ModelPerformanceandEvaluation:



## RandomForestvsBaselineModel

Model Performance

* Improvement of 4.36 % was seen inRandomForestagainstDecisionTree.

## RandomForestTunedvsBaselineandRandomForest Models

Model Performance

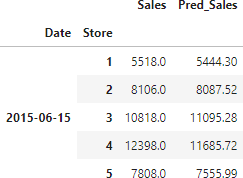
* Improvement of 4.382 % was seen inRandomForestTunedagainstDecisionTree.
* Improvement of 0.021 % was seen inRandomForestTunedagainstSimpleRandom Forest.

## Evaluation Metrics:

* Mean Absolute Error(MAE)- MAE is avery simple metric which calculates themean of absolute difference betweenactual and predicted values.
* Mean Squared Error(MSE)- Meansquared error states the mean of thesquareddifferencebetweenactualandpredicted value.
* RootMeanSquaredError(RMSE)-Itisasimplesquarerootofmeansquarederror.
* R Squared (R^2)- R2 score is a metricthat tells the performance of your model,not the loss in an absolute sense that howwell did your model perform. Hence, R2squared is also known as Coefficient ofDetermination or sometimes also knownasGoodnessoffit.It’svaluerangesfrom0 to 1. It can be negative if the model isperforming worse than the base.
* Adjusted R Squared- The disadvantageof the R2 score is while adding newfeatures in data the R2 score startsincreasing or remains constant but itnever decreases because It assumes thatwhile adding more data variance of dataincreases. Adjusted R^2 is adjusted forthis disadvantage and shows the realvalue.

# StorewiseSalesPrediction:

Here’stheheadoflatestsixweeksactualsalesvalues against the predictions which can belocated date and store wise:



# ConclusionandRecommendations:

## Conclusion:

The main objective of sales forecasting is to paintan accurate picture of expected sales. Sales teamsaimtoeitherhittheirexpectedtargetorexceedit.

When the sales forecast is accurate, operations gosmoothly and future planning for the company'sgrowthisdone efficiently.

Upon having this analysis it can be establishedthat given the dataset, the model developed isable to explain 95.5878 % of the variations and isable to predict the sales values in a good range.

Some important insights to draw from theanalysis includes:

* There were more sales on Monday,probablybecauseshopsgenerallyremain

closed on Sundays which had the lowestsales in a week. This validates thehypothesis about this feature.

* ThepositiveeffectofpromotiononCustomers and Sales is observable.
* Most stores have competition distancewithin the range of 0 to 10 kms and hadmoresalesthanstoresfaraway,probablyindicating competition in busy locationsvs remote locations.
* Store type B though being few innumber had the highest sales average.The reasons include all three kinds ofassortments specially assortment level bwhich is only available at type b storesand being open on Sundays as well.
* The outliers in the dataset showedjustifiable behavior. The outliers wereeither of store type b or had promotiongoing on which increased sales.
* Random Forest Tuned Model gave thebest results and only 0.021%improvement was seen from the basicrandom forest model which indicatesthat all the trends and patterns that couldbe captured by these models withoutoverfitting were done and maximumlevel of performance achievable by themodel was achieved.

**Recommendations:**

* More stores should be encouraged forpromotion.
* Store type B should be increased innumber.
* There's a seasonality involved, hence thestores should be encouraged to promoteand take advantage of the holidays.

## Challenges:

* The major challenge would be thecomputational time and RAM needed towork upon such a dataset in a cloudenvironment.

# References:

* Machine Learning Mastery
* GeeksforGeeks
* AnalyticsVidhyaBlogs
* TowardsDataScienceBlogs
* Built in Data Science Blogs
* Scikit-LearnOrg
* Investopedia