**Advertising Sales Channel Prediction**

**Problem Definition :**

Stopping advertising to save money is like stopping your watch to save time- Henry Ford

This is how significant advertisement is for any business. For any company advertisement is a means through which the product can reach to masses and at the same time make your product big and popular. Now as far as the channels of advertisements are considered it becomes really necessary to check which channel generates more revenue.

Advertisement sales channel project include three channels of advertisements i.e. T.V , radio and newspaper.

These three are the input variables and we have to with the help of machine learning algorithms determine sales generated i.e. our target variable from all the sales channel.

**Data Analysis:**

It is a small dataset with 200 entries i.e. 200 rows. 4 columns, 3 input variable and 1 target variable. The target variable is of continuous type therefore it is a regression type of problem.

**Data type:** The data type of all the variable is float64.

**Missing Values:**

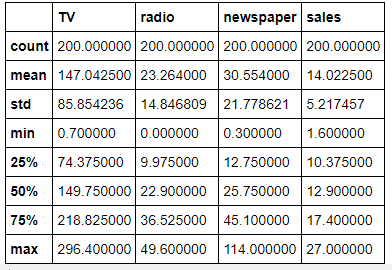
Dataset must present complete values in the features, without any missing data. Having an incomplete example makes connecting all the signals within and between features impossible. Missing values also make it difficult for the algorithm to learn during training. Too many missing values render more uncertain predictions because missing information could conceal any possible figure, consequently, the more missing values in the features , the more variable and imprecise the predictions.

Our dataset does not contain any missing values therefore no need to clean data in this aspect.

**Statistical Calculation:**

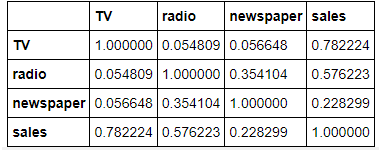
**Outlier:** Outliers are the values that are too high or too low or simply they are unusual and highly improbable. Outliers are a problem when learning from data . Because machines learn by observing examples and extracting rules and patterns from them, a weird case can prove difficult to understood and force the algorithm to reconsider what it has learned so far. A value that exists for outside the expected range will have a huge deviance from the expected value, leading the learning process to adapt to the anomaly abruptly by undervaluing the regular values (which instead produce comparatively slight deviations).

If we take into consideration statistical daya then it is observed that Max value of newspaper is very high than 75th percentile indicating presence of outlier. Also Min value of only “radio” is zero.



**Correlation :** Correlation in statistics is a measure ranging from +1 to -1 that tells how two variables relate linearly (that is, if you plot them together, they tell how lines resemble each other)

Highest correlation is between TV and sales i.e 0.78 followed by radio and sales . Heatmap shows positive correlation between all the features .



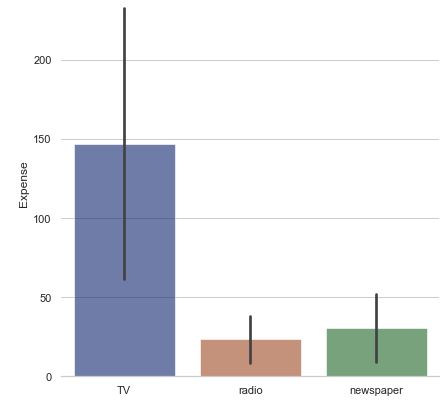
**Exploratory Data Analysis [EDA] :**

Exploratory data analysis is an approach of [analyzing](https://en.wikipedia.org/wiki/Data_analysis" \o "Data analysis) [data sets](https://en.wikipedia.org/wiki/Data_set) to summarize their main characteristics, often using [statistical graphics](https://en.wikipedia.org/wiki/Statistical_graphics) and other [data visualization](https://en.wikipedia.org/wiki/Data_visualization) methods.

**1.Catplot:**

Using pd.melt the columns were readjusted with names ‘sales’, ’Media’ and ‘Expense’

A catplot was plotted , which showed that expense of advertisement on TV is the largest followed by newspaper and lowest for radio.



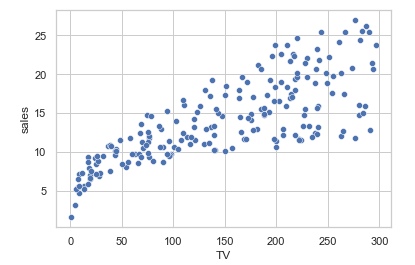
**2.Distribution Plot:**

The distribution plot of all the four columns were plotted it showed that mean and median of TV and radio are nearly equal. Both show bimodal normal distribution.

Newspaper and sales plot shows mean and median are different and away from each other. It shows skewness in its distribution.

**3. Scatter plot:**

The scatter plot with respect to sales shows that TV and sales have a linear relationship indicating increase in expense of TV also increases sales.



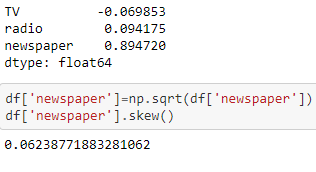
**4. Box plots:**

The box plots shows that newspaper variable has two outliers and the other the other two does not contain any outliers.

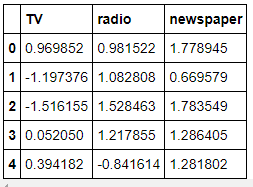
**Pre-processing:**

No matter the level of sophistication of the learning algorithm , if you don’t prepare foundation well i.e. your data –your algorithm won’t last long when tested in real data situations. According to the principle of garbage in, garbage out (GIGO for short) bad data can truly harm machine learning. Bad data consists of missing data, outliers, skewed value distribution, redundancy of information and features not well explicated.

**Skewness:** After separating independent and dependent variable it is observed that newspaper variable is little skewed. It is removed by using square root method.



**Scaling:** Standard Scaler is used to scale the data.



**Building Machine learning model:**

After the data is ordered and arranged properly now we can build our model. As this is a classification type of problem we will choose our algorithm accordingly.

1. **Finding best random state :-** Before splitting train and test data best random state is find out which comes out to be 13.

2. **Train-Test Split:** Here data is split into training data and testing data using the train- test split method. So here we have set the testing data to be 30% along with the random state to be 13.

3. **Finding Best Model:** The regression type of models used are:

**Linear Regression:** Linear regression is a [linear](https://en.wikipedia.org/wiki/Linearity) approach to modelling the relationship between a [scalar](https://en.wikipedia.org/wiki/Scalar_(mathematics)) response and one or more explanatory variables (also known as [dependent and independent variables](https://en.wikipedia.org/wiki/Dependent_and_independent_variables)). The case of one explanatory variable is called [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression); for more than one, the process is called multiple linear regression. This term is distinct from [multivariate linear regression](https://en.wikipedia.org/wiki/Multivariate_linear_regression), where multiple [correlated](https://en.wikipedia.org/wiki/Correlation_and_dependence) dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using [linear predictor functions](https://en.wikipedia.org/wiki/Linear_predictor_function) whose unknown model [parameters](https://en.wikipedia.org/wiki/Parameters) are [estimated](https://en.wikipedia.org/wiki/Estimation_theory) from the [data](https://en.wikipedia.org/wiki/Data). Such models are called [linear models](https://en.wikipedia.org/wiki/Linear_model).

**Decision Tree Regressor**: Using a sample of observations as a starting point, the algorithm retraces the rules that generated the output classes (or the numeric values when working

through a regression problem) by dividing the input matrix into smaller and

smaller partitions until the process triggers a rule for stopping. Such retracing

from particular toward general rules is typical of human inverse deduction, as

treated by logic and philosophy. In a machine learning context, such inverse reasoning

is achieved by applying a search among all the possible ways to split the

training in-sample and decide, in a greedy way, to use the split that maximizes statistical measurements on the resulting partitions

**SVR:** In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), support-vector machines (SVMs, also support-vector networks) are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-[probabilistic](https://en.wikipedia.org/wiki/Probabilistic_classification) [binary](https://en.wikipedia.org/wiki/Binary_classifier) [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier) (although methods such as [Platt scaling](https://en.wikipedia.org/wiki/Platt_scaling) exist to use SVM in a probabilistic classification setting). SVM maps training examples to points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

**Random Forest Regressor:** Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time. For classification tasks, the output of the random forest is the class selected by most trees. Random Forest is a classification (naturally multiclass) and regression algorithm that uses a

large number of decision tree models built on different sets of bootstrapped

examples and subsampled features.

As it is a regression type of problem the accuracies of different regression models are find out. Also mean error,mean squared error and root mean squared error is found out.

The r2 score for different models are:

Linear Regression: 0.9332

Decision Tree Regressor : 0.9671

SVR : 0.8632

Random Forest Regressor : 0.9771

**Cross Validation:**

A noticeable problem with the train/test set split is that you’re actually introducing

bias into your testing because you’re reducing the size of your in-sample

training data. When you split your data, you may be actually keeping some useful

examples out of training. Moreover, sometimes your data is so complex that a test

set, though apparently similar to the training set, is not really similar because

combinations of values are different (which is typical of highly dimensional datasets).

These issues add to the instability of sampling results when you don’t have

many examples. There is a chance of overfitting in the above obtained accuracy . Therefore we need to check overfitting using cross validation score. The scores obtained are as follows :

Linear Regression: 0.8871

Decision Tree Regressor : 0.9531

SVR : 0.9016

Random Forest Regressor : 0.9753

**Best Model :**

Observing the accuracy score and cross validation , we can choose our best model. As the least difference between accuracy and cross validation score is for Random Forest Regressor , we will choose it as our best model.

**Hyperparameter Tuning:**

Now we need to tune the parameters of our best model . The final source of performance derives from fine-tuning the algorithm’s *hyperparameters*,

which are the parameters that you decide before learning happens and

that aren’t learned from data. The Best parameters for hyperparameter tuning Random Forest Regressor are obtained and the model is hyper tuned which gives r2 score as 0.97716

**Concluding Remarks :**

1. Expense of advertisement on TV is the largest
2. The correlation between TV and sales is observed to be highest.
3. Random Forest regressor is our best model for prediction. .

**References:**

1.Machine Learning for dummies by John Paul Muellerand and Luca Massaron

2. Wikipedia