Retail Sales – Modelling Stage

Approach

1. Different models for Complete raw dataset (just a merge of all datasets provide) including markdown values and without any data transformation techniques other than replacing missing values of markdowns with zero.
2. Non transformed dataset as above but without markdown columns
3. Apply some data transformations like replacing dates with week numbers and apply multiple algorithms

Overall approach is to split the dataset into multiple types

1. raw dataset with all data
2. dataset with no markdowns
3. dataset with missing markdowns filled with zero
4. dataset with date converted to week number

For each type of dataset, apply multiple regression algorithms like Linear regression, Decision trees and ensemble techniques like Random Forests.

Error Metrics

As this problem to predict the Weekly Sales, which is a continuous variable, it is a regression problem. For the regression, 3 main error metrics are

1. r2 – Main purpose of choosing this metrics is to know how better is the model compared to the mean model. Its ranges from –infinity to 1 with 1 being the best model.
2. RMSE – It calculates the sum of squared differences between the predicted and actual values and then averages. This metrics is useful in the current problem context to know what's the average variation of the predicted and actual values.
3. MAE – Mean absolute error calculates the absolute differences between predicted and actual value and then averages the result. Compared to rmse, MAE will calculate only average absolute difference. This is helpful when data has outliers

Both RMSE and MAE helps in comparing between models where as r2 can be used as how good the model is from its own base model (base model is the one which predicts the average of Weekly Sales always)

Multiple models have been built taking above approach and all 3 error metrics have been calculated and listed in below table

Model No	train_r2	validation_r2	train_rmse	validation_rmse	train_mae	validation_mae
3	0.994987112	0.969933598	1641.237724	3962.769884	19.2080046	52.08788675
9	0.995503254	0.968948745	1522.282139	4009.122453	1.968604692	21.49358119
6	0.995467213	0.968809653	1528.370361	4018.091725	2.353940097	21.57505724
5	0.762911903	0.771973424	11053.52618	10864.31815	1.82716E-11	11.04672321
2	0.760843036	0.759193614	11336.23866	11214.81524	3.4509E-12	51.81069298
7	0.096562545	0.097114264	21577.19641	21618.52892	9.40143E-12	73.92400507
1	0.093351362	0.094159203	22072.26968	21751.24204	4.42412E-13	91.08942342
4	0.088333063	0.092406012	21675.24781	21674.82239	18.37963983	76.4907287
8	0.106255976	0.0771677	21461.12799	21856.02282	1.286735139	83.57151518

From the table above, Model 3, 6, 9 and 10 have the best r2, mae and rmse whereas 1, 4, 7 and 8 has the worst metrics.

Below is the list of models (Models Numbers mentioned below are same as for table above.) Ordered by datasets with poor models in the beginning.

Models 7-8

Models 7-8 are with entire dataset obtained by merging all 3 csv.
Missing markdowns are filled as zero

Model7:
DecisionTreeRegressor(criterion='mse', max_depth=7, max_features=None,
max_leaf_nodes=None, min_impurity_decrease=0.0,
min_impurity_split=None, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.0,
presort=False, random_state=None, splitter='best')
Description: Decision Tree regression Model with no data transformations. Includes all columns including markdowns. This dataset contains data prior to the introduction of markdowns
train_r2: 0.09656254543257947
train_mae: 9.401427121469734e-12
train_rmse: 21577.19641228585
validation_r2: 0.09711426429031167
validation_mae: 73.92400507043483

validation_rmse: 21618.528916645013

Model8:
Model details:
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
max_features='auto', max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=100,
n_jobs=None, oob_score=True, random_state=1, verbose=0,
warm_start=False)
Description: Random forest regression Model with no data transformations. This dataset contains data prior to the introduction of markdowns and also later
train_r2: 0.10625597624689775
train_mae: 1.2867351386963013
train_rmse: 21461.127985366205
validation_r2: 0.07716769995044315
validation_mae: 83.57151518176187
validation_rmse: 21856.022824493226

Interpretation:

Results look poor. R2 of 0.1 indicates the models is similar to the base model (estimating all values of Weekly Sales as mean).
RMSE is also very high.
Even though there is less overfitting, the model is not very useful as the predictions are poor.

Models 1-3

Models 1-3 are with dataset which contains markdown values. As markdowns are introduced towards end of year 2011, this dataset removed contains only data after markdowns are introduced.
Missing markdowns are filled as zero

Model1:
LinearRegression (copy_X=True, fit_intercept=True, n_jobs=None, normalize=True)
Description: Linear regression Model with no data transformations. Includes all columns
train_r2: 0.09335136190344195
train_mae: 4.4241154264012437e-13
train_rmse: 22072.269683556686
validation_r2: 0.09415920290341961
validation_mae: 91.08942341878617
validation_rmse: 21751.242044027294

Model2:
Model details:
DecisionTreeRegressor(criterion='mse', max_depth=7, max_features=None,

max_leaf_nodes=None, min_impurity_decrease=0.0,
min_impurity_split=None, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.0,
presort=False, random_state=None, splitter='best')
Description: Decision Tree regression Model with no data transformations. Includes all columns
train_r2: 0.7608430604905764
train_mae: 3.4509025873090033e-12
train_rmse: 11336.238663526812
validation_r2: 0.7591936136791231
validation_mae: 51.81069297940961
validation_rmse: 11214.815241125414

Model 3:
Model details:
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
max_features='auto', max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=100,
n_jobs=None, oob_score=True, random_state=1, verbose=0,
warm_start=False)
Description: Random forest regression Model with no data transformations. Includes all columns
train_r2: 0.9949871119356408
train_mae: 19.208004602480617
train_rmse: 1641.2377235868435
validation_r2: 0.969933597866702
validation_mae: 52.08788675189235
validation_rmse: 3962.7698842137193

Interpretation:

Linear regression performs poorly then the other models. This might be because the assumptions of linear regression like correlation between variables might exists.
Decision tree performs much better than the linear regression.
TO add more randomness to the decision trees, random forests is also tried and the metrics looks much better than the other models.
R2 for random forest 0.99 for training and 0.96 for the test set signifies that the model is much better than the base model.
MAE 51 for validation set looks good considering the scale of Weekly Sales is from -200 to over 100000. Despite of Weekly Sales spread over such a large scale, MAE of 51 looks good.
RMSE 3962 is higher than the MAE may be because of the outliers in the dataset. RMSE is susceptible to outliers because it tries to square the difference. If for some noisy row, the prediction goes wrong, squaring this difference adds more weight on the error. Hence RMSE is high.

The random forest is best candidate for the above dataset as it has good predictions with low errors and also it generalises well. Result above proves that there is very minimal overfitting.

Models 4-6

Models 4-6 are for the dataset with no markdown data. There might be scenario where the markdowns may not be present for future sales, this model helps in predicting those data

Model4:
Model details:
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=True)
Description: Linear regression Model with no data transformations. Includes all columns except markdowns
train_r2: 0.08833306322917589
train_mae: 18.379639828498238
train_rmse: 21675.247811126523
validation_r2: 0.092406011546733
validation_mae: 76.49072870460421
validation_rmse: 21674.82239013148

Model5:

```
DecisionTreeRegressor(criterion='mse', max_depth=7, max_features=None,
                      max_leaf_nodes=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      presort=False, random_state=None, splitter='best')
```

Description: Decision Tree regression Model with no data transformations. Includes all columns except markdowns
train_r2: 0.7629119032567171
train_mae: 1.827161631005865e-11
train_rmse: 11053.526181981952
validation_r2: 0.7719734240466052
validation_mae: 11.046723206674706
validation_rmse: 10864.318152665051

Model6:

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
                      max_features='auto', max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, n_estimators=100,
                      n_jobs=None, oob_score=True, random_state=1, verbose=0,
                      warm_start=False)
```

Description: Random forest regression Model with no data transformations. Includes all columns except markdowns
train_r2: 0.9954672133720925
train_mae: 2.3539400974334352
train_rmse: 1528.3703612895772
validation_r2: 0.9688096528152962
validation_mae: 21.575057240790354
validation_rmse: 4018.091725050686

Interpretation:

Linear regression performs poorly then the other models. This might be because the assumptions of linear regression like correlation between variables might exists.
Decision tree performs much better than the linear regression.
TO add more randomness to the decision trees, random forests is also tried and the metrics looks much better than the other models.
R2 for random forest 0.99 for training and 0.96 for the test set signifies that the model is much better than the base model.
MAE 21 for validation set looks good considering the scale of Weekly Sales is from -200 to over 100000. Despite of Weekly Sales spread over such a large scale, MAE of 51 looks good.
RMSE 4108 is higher than the MAE may be because of the outliers in the dataset. RMSE is susceptible to outliers because it tries to square the difference. If for some noisy row, the prediction goes wrong, squaring this difference adds more weight on the error. Hence RMSE is high.

The random forest is best candidate for the above dataset as it has good predictions with low errors and also it generalises well. Result above proves that there is very minimal overfitting.

If the purpose of the problem is to predict the Week Sales alone, this dataset with Random forest has even low error scores compared to the dataset with markdowns. This also proves that, markdowns are not a important to predict Weekly Sales. This also indicates that, spikes of sales were not because of the promotions but they are the general trend which is seen at the end of year

Business Insights:

1. Model built from whole dataset replacing missing values of markdowns with zero performs very poor. This indicates that the data in the raw form is not suitable for modelling. This could be because of large number of markdowns marked as zero and that might have bad effects on the Weeky sales.
2. Models build from dataset which contains markdowns (removed data prior to introduction of markdowns in year 2011) make good predictions.
3. Models build from dataset containing no markdowns also make good predictions.

Above points indicates that markdowns are not very useful in predicting the Weekly Sales. If the purpose of the problem is to predict the Week Sales alone, this dataset with Random forest has even low error scores compared to the dataset with markdowns. This also proves that, markdowns are not a important to predict Weekly Sales. This also indicates that, spikes

of sales were not because of the promotions but they are the general trend which is seen at the end of year.

Final Thoughts:
Approach of splitting the dataset into multiple families (like dataset with no markdown, dataset after markdown introduction, raw dataset, etc.) is useful in maintaining the predictive models for the company. As there is no guarantee that the data will be in the same character and form in the future (like markdowns being dropped), having different models suiting different kinds of data helps the company in some savings in terms both time and cost.

As the scope of this submission is limited to the modelling and choosing the best model, more analysis about each variable impact on the model to be made in the final project report in coming weeks.