20092603

STAT0030: ICA2

Exploratory Data Analysis

Firstly, the grocery.csv dataset was loaded in R. Based on the nature of the variables some of them needed to be transformed into different types. "STORE_NUM","UPC", "MANUFACTURER" should be factors while "DISPLAY","FEATURE","TRP_ONLY" should be logical/boolean.

Since UNITS is the response variable, which takes non-negative integer values, there are three more numerical values: PRICE, BASE_PRICE and WEEK_END_DATE We get some summary statistics for numerical variables. By just looking at the means and min/max values it is clear that UNITS is heavily skewed.

	nbr.na	min	max	range	median	mean	var	coef.var
BASE_PRICE	0	2.8	8.9	6.1	6.4	5.7	2.0	0.3
PRICE	0	1.8	8.9	7.1	5.5	5.3	2.0	0.3
WEEK_END_DATE	0	39827.0	40912.0	1085.0	40401.0	40393.0	96182.7	0.0
UNITS	0	1.0	153.0	152.0	8.0	12.3	187.8	1.1

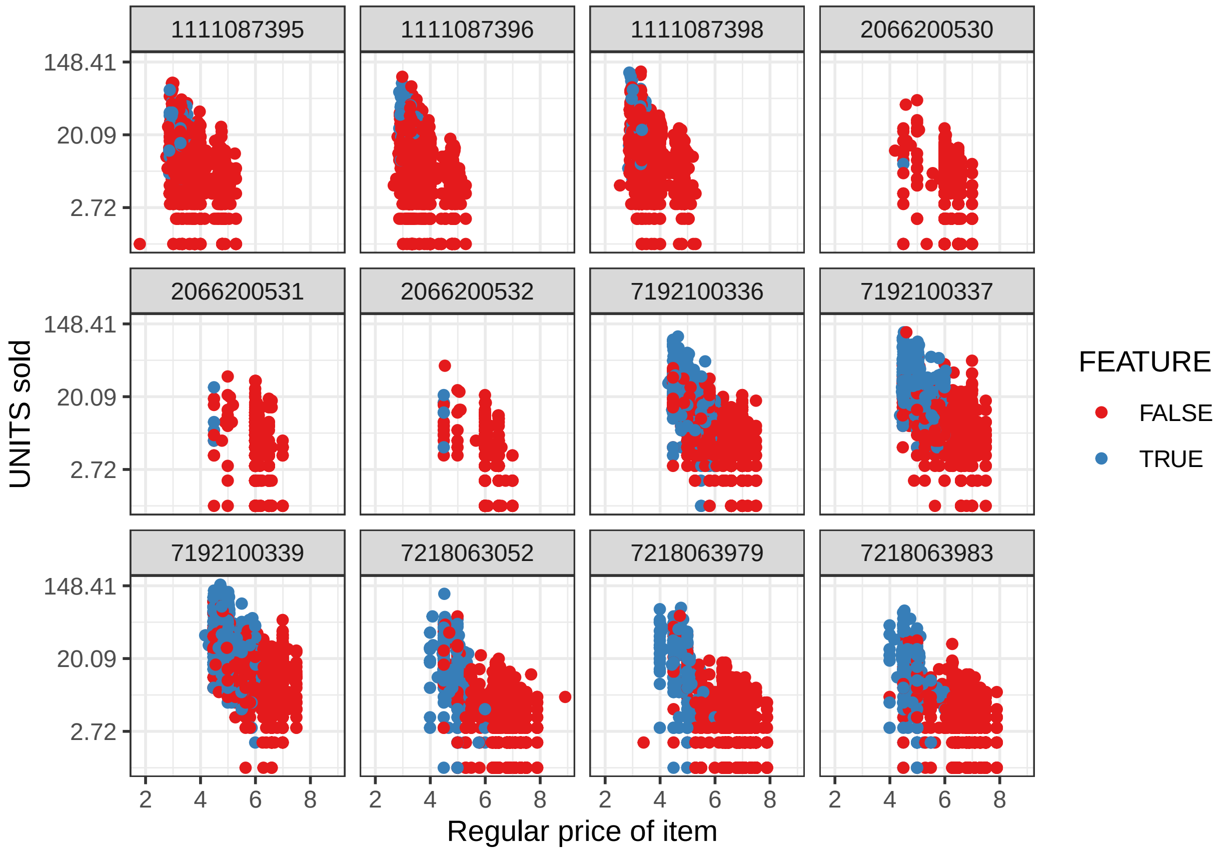
Based on the scatterplots, correlation coefficients and density plots for the numerical variables the distribution of BASE_PRICE seems to be bimodal, perhaps a mixture of 2 or 3 different sub distributions. The distribution of PRICE seems to have 3 peaks while the distribution of UNITS seems to be Gamma with longer tail, perhaps log-normal. There is some clear linear relationship between PRICE and BASE PRICE. There is also a weak linear relationship between PRICE and UNITS.

By plotting the PRICE against UNITS for every manufacturer and indicating the items on DISPLAY with different color it becomes clear that items on display amount more sales, especially for prices on the lower end of the PRICE distribution for each MANUFACTURER.

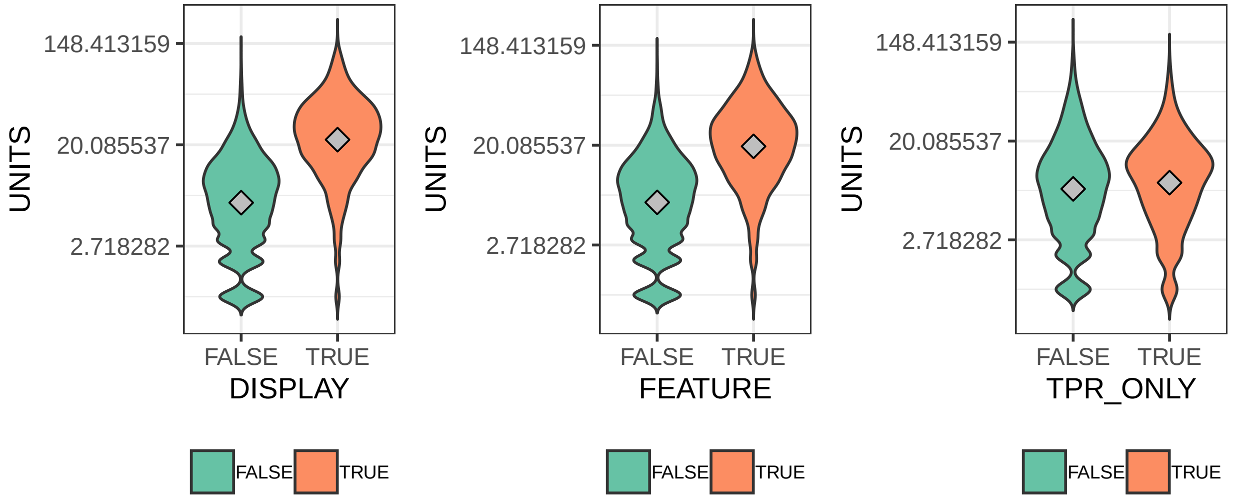
In the following plot, items that were featured on in-store leaflet amount more sales, especially for prices on the lower end of the PRICE distribution for each product. The negative linear relationship between the log of UNITS and PRICE becomes clearer. From the violin plots of UNITS vs the three boolean variables it is evident that there is a difference in the distribution for each level of DISPLAY and FEATURE, but not a significant one for TPR_ONLY, which suggests that it does not carry much information about the variable UNITS. Also,the distribution of UNITS is heavily skewed. By applying log transformation the distribution becomes more symmetrical.

Scatterplot of Price vs Units sold per product (Lin−Log scale)

Different colors based on whether product was in in−store leaflet



Violin plots of Units sold (log scale) vs all of the boolean variables



Considerarations for repsonse variable distribution for GLM models

The distribution that seems to fit the response variable (UNITS) better is the Log-Normal. However, since it is not a part of the exponential family it cannot be used to fit a GLM model. The Gamma distribution could potentially be a candidate. On the other hand, if we consider the UNITS variable as a count variable, since it represents the amount of sales, it would be reasonable to try a Poisson GLM regression model. However, as it is suggested by the log-likelihhod of the best fitted Poisson distribution the fit is rather poor.

By further inspection of the distribution and the GLM model (next section), there's evidence of overdispersion (overidispersion parameter phi is estimated to be ~5 instead of 1 which is assumed and held constant by Poisson regression). Possible solutions to this problem could be fitting a quasipoisson regression model or a Negative Binomial GLM model. It also seems that that the Gamma or Negative Binomial might be fit better since the likelihoods of their distributions are also higher than that of the Poisson distribution.

	Likelihood	parameter1	parameter2
LogNormal	-34600.88	2.042	1.000
Gamma	-34995.87	1.205	0.098
Exponential	-35099.43	0.081	NA
NegBinomial	-35301.74	1.341	12.322
ChiSquared	-45805.81	8.687	NA
Poisson	-73272.10	12.322	NA

Linear Regression Models

Based on the EDA for the grocery dataset the variables STORE_NUM, TPR_ONLY and MANUFACTURER seem to not carry significant amount of information regarding the response variable. So for simplicity of the models they will not be included. Also, BASE_PRICE is highly correlated with PRICE and to avoid violating the collinearity assumption for the covariates it will also be excluded from the models. The Linear Model with the covariates PRICE, WEEK_END_DATE, UPC, FEATURE, DISPLAY had a very low R^2 (45) and there were severe violations of the assumptions of homoscedasticity and normality of the residuals. A cure for that seemed to be the Log-transformation of the response variable (UNITS). This also improved the R^2 (49.4).

The data were splitted in 10 folds and the first nine were used as a training set. The RMSE of the predictions made by the Log transformed Linear model was 9.872.

Generalised Linear Regression Models

Poisson Regression

If we treat the response variable as a count variable, the first thing that comes in mind is Poisson regression. This was the first attempt, Poisson GLM with Log-link but the residual deviance was rather high, heteroscedasticity was present based on residual vs fitted plot and the normality assumption was violated based on residual's Q-Q plot.

The Poisson regression model keeps the dispersion parameter fixed at 1, implying equality of the mean and the variance. The variance and mean ratio for UNITS is 15.24 >> 1 which implies overdispersion in our data. An estimation of the dispersion parameter as Residual deviance divided by the residual degrees of freedom yielded $\hat{\phi} = 4.72$. The following are some ways to account for overdispersion in the data.

Quasipoisson, Negative Binomial, Quasi Regression

A quasipoisson regression model with Log-link was fitted. The dispersion parameter was estimated as 5.25. The residual deviance is lower than the Poisson model, so is the QAIC (although I am not depending on QAIC too much since it is not widely accepted). The assumptions that were violated in the Poisson case are still not cured. Negative Binomial regression models were also considered. The one with the log-link gave significantly lower residuals deviance (about 1/4 of the Poisson deviance). It also managed to cure the heteroscedasticity to a certain point. The residual's normality is still violated on both tails. Several combinations of variance and link functions were considered for the Quasi regression models, such as the variance being a function of the mean or the second and third power of the mean and link functions such as the logarithm, square root and the identity function. Some interesting results were in the case of the variance being a equal to the square of the mean with log link function in which the deviance was much lower than all of the previous models, the residuals were almost normally distributed but the residual's variance was not constant. Another model was with the variance being equal to the mean, a log link function and a log transformation on the response. This resulted in the lowest deviance so far, with residuals normally distributed and with their variance not being far from constant.

Gamma Regression: Since the UNITS variable takes many different values on the $(0, \infty)$ interval it can be treated as a continuous variable. The Gamma distribution is part of the exponential family of distributions and so we can fit a GLM model with the assumption that the response is following a Gamma distribution, since it seemed from the previous analysis that it fits the data well. The residuals deviance was similar to the quasi regression models, although the diagnostic plots were similar to the Negative Binomial model.

Predictions: In terms of predictions on the 10 test set, the Poisson and Quasipoisson with loh links performed similarly, with RMSEs 9.6979 with the Negative Binomial regression being slightly worse 9.72424 and then the Gamma and the quasi regression with var=mu^2 with 9.72716. The worst performing models were the linea regression and the quasi with transformed response.

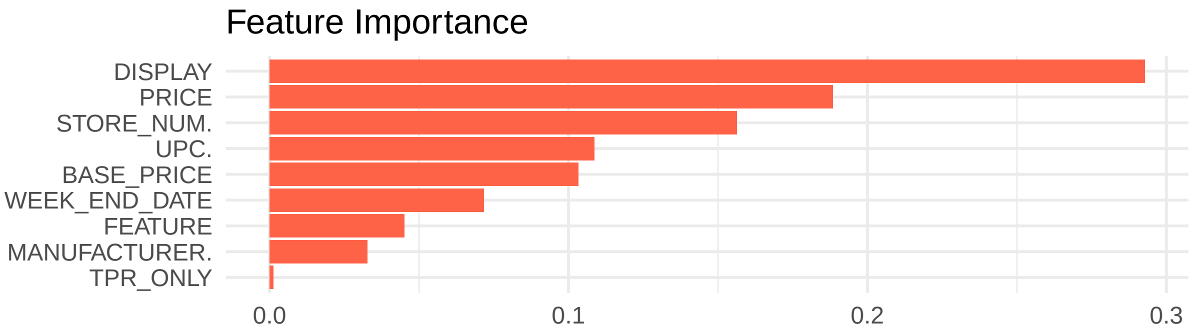
Modern Regression models

Regression Tree and Random Forest

A Regression Tree model was fitted on the training set with the same inputs as the GLM models. Using 10-fold Cross Validation to determine optimal size led us to keep the whole tree. The RMSE on the 10th fold was 9.72976. As for the Random Forest, a plot of the importance of the features confirmed that the covariates that were excluded so far had the lowest increase in MSE when they are removed from the model. A custom grid search using 5 fold CV was used in order to determine the optimal values for number of trees (ntree for values 300,500,1000) and the number of variables available for splitting at each tree node (mtry for values 2,3,4). The optimal parameters were $ntree = 1000$ and $mtry = 4$. The RMSE on the 10th fold was 8.74104.

XGBoost

For this model, the categorical and boolean variables were transformed to dummy variables (one-hot encoding). Then a plot of the features' importance confirmed our findings about the variables in the dataset. The optimal hyperparameter values for nrounds and maximum depth were assessed through 5-fold CV and were found to be 313 and 5 respectively. The RMSE of the predictions on the 10th fold was 6.53811, which is significantly better to all of the previous models.



Comparison of models

The models that were chosen was the Quasipoisson with log link function from the GLM family and the XGBoost from the modern regression models. The latter's main advantage is the high prediction quality, even though it is rather complex to explain, visualise or get the intuition when making predictions. The GLM is much more straightforward, the log link suggests that logarithm of the mean of the response if a function of the covariates. The Quasipoisson model seemed to perform better than the rest of the models and equally well as the Poisson one. But for the latter one, the assumptions were violated to a larger degree. As for the XGBoost model, it simply outperformed all the other models by a considerable margin. The mean RMSE for the 10 folds was 9.8133 for the Quasipoisson and 6.7576 for the XGBoost.

A paired t-test was perform for the RMSEs of the two models for each of the 10 folds. The t-test ha a p-value of the order 10^{-8} which suggests that the mean is significantly different than zero, which mean that the RMSE is different (much lower for the XGBoost).

Final model

The final model for the task of prediction of UNITS variable or the sales of grocery products was an XGBoost model, with weak learners (trees) of maximum depth 5 and a maximum number of 313 boosting iterations. The number of features were 96, most of which are the dummified version of categorical variables. The eta parameter could also be optimised further, since the default value of 0.3 was used. The features of the model were PRICE, BASE_PRICE, WEEK_END_DATE, STORE_NUM, UPC, DISPLAY and FEATURE. Based on the importance of the features of the optimised XGBoost model and with respect to Gain, DISPLAY, PRICE and STORE_NUM were the most important features. Hence, we may presume that the sales of a product are dependent on whether it was on display in the store. Other than that, the different stores seem to affect the sales as well as the initial price of the products.

This final model estimates for the product with UPC = 7192100337, during WEEK END DATE = 39995 at STORE NUM = 8263 that by decreasing the PRICE by 10 and keeping all of the other covariates constant that there would be 11.70641 UNITS sold. Of course, we need an integer value so we would probably go for 12 UNITS. That's a decrease of 1 UNITS from the original PRICE.

Appendix

In most of the residual vs fitted plots for lm and glm models there appear to be parallel lines. After some brief research on the topic, it appears to be natural for data with integer values. References: (https://www.tandfonline.com/doi/abs/10.1080/00031305.1988.10475569#) (http://www.mensurationists.com/docs/conf2016/Wang_Mingliang.pdf)