# Set2_Tsk4

A brief note on the approach to this subtask: since the task description requires the selection of three variables based on economic judgement and the interpretation of the logistic regression model using these variables, this is how I will proceed.

An alternative approach would be to fit a model incorporating all regressors and, if appropriate, compare it to a Lasso model. This approach would essentially treat the problem as a classic forecasting task. Although key metrics are initially easier to understand and interpret, using all available data (i.e. regressors) in this way could improve predictive performance.

## Selection of three predictors

The idea is now to select three predictors that have as little conceptual overlap as possible. This will prevent redundant variables from being included in the model, which could otherwise negatively affect its interpretability (cf. multicollinearity).

### 1. Trend in Monetary value of sales: *sales_trend_3m*

The total sales values can be compared automatically across different products and categories. It is also intuitive to assume that regions or stores with persistently low sales are potential candidates for closure. However, it is important to note that absolute sales figures are not a suitable metric, since they can vary substantially between stores. For example, a small but profitable store may generate less revenue than a large but unprofitable one.

For this reason, the three-month trend is considered here: *sales_trend_3m*, which is the relative change in average sales between the first three months and the subsequent three months of the pre-closure period. Changes at the monthly level are less robust since short-term fluctuations are to be expected. For instance, sales may temporarily decrease for non-critical reasons, such as an abundance of public holidays or supply bottlenecks, or temporarily increase due to pre-Christmas shopping.

**H1:** *A decline in sales over several months is an early warning sign of structural weakness in demand, which increases the risk of closure.*

### 2. Percentage of sales generated online: *pct_online_sales*

If a large proportion of customers shop online, this results in lower margins for the offline business. This can cause a store to become unprofitable in the long term.

**H2:** *A high proportion of sales made online indicates a structural shift in demand from physical stores to digital channels. This reduces the profitability of local stores and increases their risk of closure.*

### 3. Percentage of discounts offered: *pct_discounts*

Using discounts to stimulate demand may indicate that a store has weak pricing power and is therefore at risk of closure. However, it is important to note that persistently high levels of discounting may also be a

deliberate strategic choice. This variable is not reported at the monthly level, so it cannot be transformed into a relative change measure. It is therefore used in its original (untransformed) form.

**H3:** *A high proportion of discounts suggests weak local demand and limited pricing power. This indicates insufficient profitability and an increased risk of closure.*

## Model analysis

To prevent information leakage, i.e. the independent variables being potentially influenced by store closures, the dataset is first filtered to include only the months prior to the closures. Next, a variable is constructed to capture the change in total sales over three months, and all other non-selected regressors are removed from the dataset. After the data transformation, since all three variables are available exclusively at the county level, duplicate observations are dropped using *'distinct()'*.

```
paste0("Proportion of closed stores: ",
       length(df_pre__treat_treated$treat)/length(store_df_pre_treat_3selected$treat))
```

```
## [1] "Proportion of closed stores: 0.5"
```

Next, a logistic regression is fitted for each of the eight possible combinations, as well as for the null model. Additionally, the in-sample classification accuracy with a threshold of 0.5 is computed for each model. A likelihood-ratio test is also performed for each model. The logistic regression models are stored in a dataset that is sorted in descending order by AIC value.

```
knitr::kable(logit_model_results_overview)
```

| Formula | AIC | Accuracy | LR statistic | p-value |
|---|---|---|---|---|
| treat ~ sales_trend_3m | 92.7278 | 0.5152 | 2.7677 | 0.0962 |
| treat ~ pct_discounts | 93.1854 | 0.6061 | 2.3100 | 0.1285 |
| treat ~ 1 | 93.4954 | 0.5000 | 0.0000 | 1.0000 |
| treat ~ pct_discounts + sales_trend_3m | 94.0900 | 0.5606 | 3.4055 | 0.1822 |
| treat ~ pct_online_sales + sales_trend_3m | 94.5488 | 0.5455 | 2.9466 | 0.2292 |
| treat ~ pct_online_sales + pct_discounts | 95.0337 | 0.6061 | 2.4618 | 0.2920 |
| treat ~ pct_online_sales | 95.0889 | 0.5000 | 0.4065 | 0.5237 |
| treat ~ pct_online_sales + pct_discounts + sales_trend_3m | 95.9833 | 0.5455 | 3.5121 | 0.3192 |

In terms of AIC, all models are very close to each other (with a difference of only 3.3 between the best and worst model). The best model uses *sales_trend_3m* as the sole regressor, whereas the worst model incorporates all three of the aforementioned economically motivated regressors. 50% of the stores in the dataset are closed (see above). Consequently, this is the automatic baseline in-sample accuracy of the null model. None of the other logistic regression models achieves substantially higher accuracy than this baseline (the best value is 60%). The null hypothesis of the LR test cannot be rejected at the 5% significance level for any of the models. Only the AIC-best model is statistically significant at the 10% level. Therefore, none of the additional parameters in the unrestricted models leads to a statistically significant improvement in the likelihood of observing the data.

**Best-performing model according to AIC: *sales_trend_3m* only**

```r
summary(logit_model_results$model[[1]])
```

```
##
## Call:
## glm(formula = f, family = binomial(link = "logit"), data = store_df_pre_treat_3selected)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.0339     0.6859  -1.507    0.132
## sales_trend_3m   1.8163     1.1158   1.628    0.104
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 91.495  on 65  degrees of freedom
## Residual deviance: 88.728  on 64  degrees of freedom
## AIC: 92.728
##
## Number of Fisher Scoring iterations: 4
```

With a p-value of 0.104 for sales_trend_3m, we cannot reject the null hypothesis that an increase in the three-month sales trend has no effect on the probability of store closure. It should also be noted that multicollinearity is not an issue in models with a single regressor.

For the sake of completeness, the following interpretation of this coefficient is provided; however, due to the low level of statistical significance, it should be treated with caution. A relative increase in average sales over the three-month period is associated with higher log-odds of store closure. Specifically, a 100% increase in the three-month sales trend (i.e. doubling sales) increases the log-odds of store closure by 1.8163. According to the model, therefore, a positive sales growth trend increases the risk of closure and contradicts hypothesis H1.

### Worst-performing model according to AIC: all three regressors

```r
summary(logit_model_results$model[[nrow(logit_model_results)]])
```

```
##
## Call:
## glm(formula = f, family = binomial(link = "logit"), data = store_df_pre_treat_3selected)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.8497     3.6784   0.231    0.817
## pct_online_sales   0.3707     1.1351   0.327    0.744
## pct_discounts     -0.2842     0.3809  -0.746    0.456
## sales_trend_3m     1.2961     1.2748   1.017    0.309
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 91.495  on 65  degrees of freedom
## Residual deviance: 87.983  on 62  degrees of freedom
## AIC: 95.983
##
## Number of Fisher Scoring iterations: 4
```

In the model incorporating all three regressors, none of the predictors are statistically significant (all p-values greater than 0.30). Due to the lack of evidence, the coefficients are not interpreted further or examined for multicollinearity.

## Conclusion

The results of the logistic regression analysis show that the three selected predictors do not have any statistically significant explanatory power when it comes to predicting whether a county experiences a store closure. A larger dataset would be interesting to work with, as it would allow for a more reliable model fit and greater statistical power. Moreover, with a larger sample, the predictive performance on unseen data could be evaluated using a train–test split.