Set2_Tsk4

Vorab zum Vorgehen in dieser Teilaufgabe: Da in der Aufgabenbeschreibung explizit gefordert ist, basierend auf der eigenen ökonomischen Einschätzung drei Variablen auszuwählen und das Logistische Regressionsmodell basierend auf diesen Variablen zu interpretieren, wird dies auch so durchgeführt. Wäre die Aufgabenstellung offener gestaltet, hätte ich ein Modell mit allen Regressoren gefitted und gegebenenfalls mit einem Lasso-Modell verglichen. Dieses Vorgehen wäre deshalb von Vorteil, da es sich bei diesem Problem klassisch um einen Forecasting Task handelt. Zwar sind Key metrics zunächst leichter zu verstehen und zu interpretieren, warum sollte man aber vorhandene Daten (i.e. Regressoren) nicht nutzen, wenn man sie hat? Wichtig zu erwähnen ist hier, dass Multikolinearität kein Problem für Forecasting darstellt. Es könnten alle vorhandenen Regressoren verwendet werden (sofern die Anzahl der Beobachtungen >= Anzahl der Regressoren) ist.

Auswahl von drei Prädiktoren

Die Idee ist es nun, drei Prädiktoren so auszuwäheln, dass sie möglichst keine inhaltlichen Überschneidungen haben, was gegebenenfalls zu redundanten Variablen im Modell führen und die Interpretierbarkeit negativ beeinflussen könnte (Vgl. einleitende Worte bzgl. Multikolinearität).

Drei monats trends von monetary value of sales

Die total sales values sind automatisch über verschiedene Produkte/Kategorien hinweg vergleichbar. Darüber hinaus ist es intuitiv anzunehmen, dass Regionen bzw. Stores mit anhaltend niedrigem Umsatz als Kandidaten für eine Schließung in Frage kommen. Wichtig ist jedoch zu beachten, dass die absolute Höhe der sales sich nicht als metric eignet, da diese zwischen stores stark schwanken kann. Ein kleiner profitabler Store kann weniger Umsatz haben als ein unprofitabler großer Store. Deshalb wird hier der 3 monats trends betrachtet. Die Veränderung auf monatlicher ebene ist weniger robust, da kurzfristige schwankungen zu erwarten sind. Zum Beispiel können aus unbedenklichen gründen die sales kurzfristig einbrechen (e.g. viele Feiertage oder Lieferengpasse) oder sehr stark sein (e.g. Vorweihnachtsgeschäft).

H1: Ein anhaltender Rückgang der Sales über mehrere Monate ist ein Frühwarnsignal für strukturelle Nachfrageschwäche und erhöht das Schließungsrisiko.

Percentage of sales generated online

Wenn ein hoher Anteil an Kunden online einkauft, führt das zu geringeren margen im offline geschäft. Dies kann dazu führen, dass ein Store strukturell unprofitabel wird.

H2: Êin hoher Online-Anteil an Verkäufen deutet auf eine strukturelle Verschiebung der Nachfrage vom stationären zum digitalen Kanal hin, was die Profitabilität des lokalen Stores verringert und damit das Schließungsrisiko erhöht.

Percentage of discounts offered

Falls Rabatte dazu eingesetzt werden, Nachfrage zu generieren, kann das ein Hinweis darauf sein, dass ein Store eine geringes Preissetzungsvermögen hat und von Schließung gefährdet ist. Ein wichtiger Punkt hierbei

ist jedoch, dass ein dauerhaft hohes Discount-Niveau auch eine bewusste Strategie sein kann. Allerdings wird diese variable nicht auf dem monatlichen Level reported und kann daher nicht zu relative change umgewandelt werden und wird unverändert verwendet.

H3: Ein hoher Anteil an discounts deutet auf eine lokal schwache Nachfrage und mangelnde Preissetzungsmacht hin. Dies ist ein Anzeichen für mangelnde Profitabilität und ein erhöhtes Schließungsrisiko.

Im weiteren Verlauf der Aufgabe werden logistische Regressionen für jede mögliche Kombination aus diesen drei Prädiktoren gefitted und über das AIC und den Likelihood-Ratio-Test verglichen.

Modellanalyse

Um information leakage (independent variablen werden durch die store closure potenziell beeinflusst) zu verhindern, wird der Datensatz zunächst auf die Monate vor den Schließungen gefiltert. Anschließend wird die Variable für die Veränderung der total sales auf drei-monats-basis erstellt und der Datensatz von allen anderen nicht ausgewählten regressoren bereinigt. Da alle der drei Variablen nach der Datentransformation ausschließlich auf county level vorliegen, werden die mehrfach vorliegenden Beobachtungen mit 'distinct()' verworfen.

```
library(readr)
library(dplyr)
```

```
##
## Attache Paket: 'dplyr'
```

```
## Die folgenden Objekte sind maskiert von 'package:stats':
##
##     filter, lag
```

```
## Die folgenden Objekte sind maskiert von 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
library(purrr)
library(broom)
library(knitr)

store_df <- read.csv("../data/StoreData.csv")

store_df_pre_treat <- store_df %>%
  filter(month <= 6)


# create relative three month change predictor for sales_value_total
store_df_pre_treat <- store_df_pre_treat %>%
  group_by(county_id) %>%
  mutate(
    sales_early = mean(sales_value_total[month %in% 1:3], na.rm = TRUE),
    sales_late  = mean(sales_value_total[month %in% 4:6], na.rm = TRUE)
  ) %>%
  ungroup() %>%
```

```
  mutate(
    sales_trend_3m = (sales_late - sales_early) / sales_early
  ) %>%
  select(-sales_late, -sales_early)

store_df_pre_treat_3selected <- store_df_pre_treat %>%
  select(county_id, month, treat, sales_trend_3m, pct_online_sales, pct_discounts)

store_df_pre_treat_3selected <- store_df_pre_treat_3selected %>%
  group_by(county_id, treat) %>%
  distinct(pct_online_sales, pct_discounts, sales_trend_3m)
```

Im hier nicht gezeigten Code-Block, wird schließlich für jede der 8 möglichen Kombinationen und das Nullmodel eine logistische Regression gefitted. Zusätzlich wird jeweils die in-sample classifiaction accuracy mit einer Schwelle von 0.5 berechnet und der Likelihood-Ratio Test durchgeführt. Die Modelle werden in einem Datensatz gespeichert, der absteigend nach der Größe des AIC sortiert ist.

```
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