Einführung in Data Science & maschinelles Lernen

Abschlusspräsentation

Gruppe 1: Manuel Böhm, Sophia Gentner, Teresa Giek & Robert Mühldorfer

Erstellte Variablen

UNIVERSITÄT MANNHEIM
Fakultät für Betriebswirtschaftslehre

Gästeübernachtungen je Monat

- Daten für ganz Schleswig-Holstein wurden entfernt, da Kiel selbst weniger touristisch geprägt ist als sein Umland.

Seeschifffahrt je Monat

- Daten für ganz Schleswig-Holstein wurden zunächst beibehalten, da Kiel als wichtiger Fähr- und Kreuzfahrthafen eine zentrale Rolle spielt.

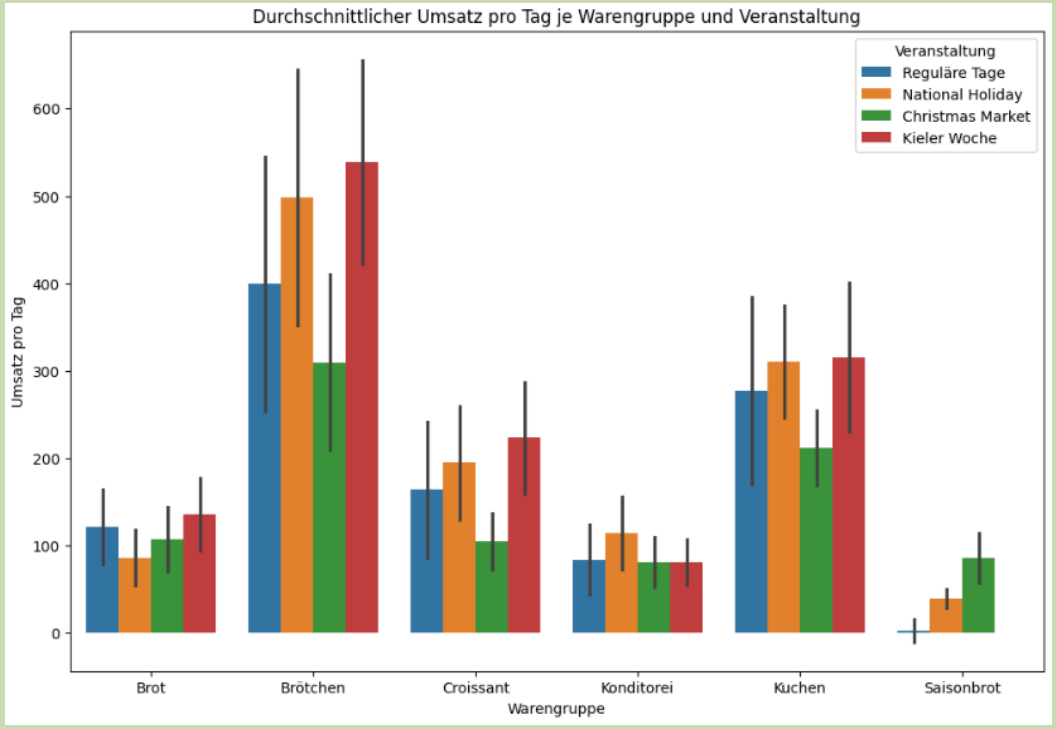
Events

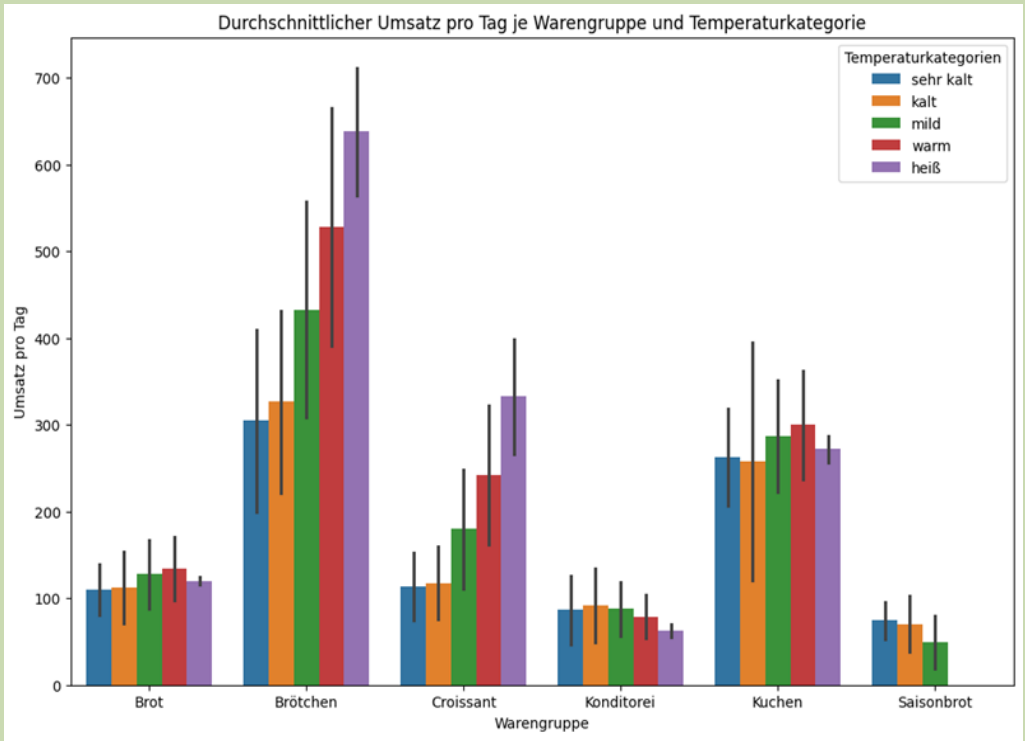
- Berücksichtigung besonderer Ereignisse wie Weihnachtsmärkte und Feiertage.

Temperaturkategorien

- Klassifizierung in "Sehr kalt", „Kalt“, „Mild“, „Warm“ und „Heiß“.

Balkendiagramme mit Konfidenzintervallen





Optimierung des linearen Modells

UNIVERSITÄT MANNHEIM
Fakultät für Betriebswirtschaftslehre

Lineares Modell

= 38.02 + 23.41 Brot + 312.58 Brötchen + 66.17 Croissant – 9.46 Konditorei + 182.94 Kuchen + 20.88 Monat_2 + 11.56 Monat_3 + 30.45 Monat_4 + 37.40 Monat_5 + 42.71 Monat_6 + 94.09 Monat_7 + 122.02 Monat_8 + 44.20 Monat_9 + 42.79 Monat_10 + 13.62 Monat_11 + 59.72 Monat_12 + 53.30 national_holiday – 47.58 christmas_market + 35.97 KielerWoche + 2.20 temp_bins_kalt + 2.84 temp_bins_mild + 7.91 temp_bins_warm + 28.50 temp_bins_heiß – 4.50 Wochentag_Di – 4.16 Wochentag_Mi + 1.46 Wochentag_Do + 3.75 Wochentag_Fr + 48.90 Wochentag_Sa + 54.99 Wochentag_So

```
mod = smf.ols('Umsatz ~ Brot + Broetchen +
Croissant + Konditorei + Kuchen + Monat_2 +
Monat_3 + Monat_4 + Monat_5 + Monat_6 +
Monat_7 + Monat_8 + Monat_9 + Monat_10 +
Monat_11 + Monat_12 + national_holiday +
christmas_market + KielerWoche +
temp_bins_kalt + temp_bins_mild +
temp_bins_warm + temp_bins_heiß +
Wochentag_Di + Wochentag_Mi + Wochentag_Do +
Wochentag_Fr + Wochentag_Sa + Wochentag_So',
data=df_training).fit()
# Output the summary of the fitted model
print(mod.summary())
```

```
R-squared:                       0.741
Adj. R-squared:                  0.740
F-statistic:                     737.4
Prob (F-statistic):               0.00
Log-Likelihood:                -42962.
AIC:                         8.598e+04
BIC:                         8.619e+04
```

Umgang mit fehlenden Werten

UNIVERSITÄT MANNHEIM
Fakultät für Betriebswirtschaftslehre

Listwise Deletion

Einfügung fehlender Umsatz-Werte

- Jede Warengruppe an jedem Tag (z.B. Saisonbrot außerhalb der Saison)
  - Umsatz von 0, wenn es eine Warengruppe ohne fehlenden Wert am gleichen Tag gibt
- Umsatz von 0 an plausiblen Schließzeiten (z.B. 25.12.)

Imputation: k-Nearest Neighbours (NN = 5)

Variable	Fehlenden Werte	
	n	prozentual
Umsatz	402	4,13 %
Temp_bins	29	0,30 %

Listwise Deletion

- **Modellgüte des neuronalen Netzes sank → Entscheidung für listwise deletion**

Optimierung des neuronalen Netzes

UNIVERSITÄT MANNHEIM
Fakultät für Betriebswirtschaftslehre

- Source Code zur Definition des neuronalen Netzes

```
# Model Definition
model = Sequential([
    InputLayer(shape=(training_features.shape[1], )),
    BatchNormalization(),
    Dense(32, activation='relu'),
    Dropout(0.1),
    Dense(16, activation='relu'),
    Dense(1, activation='linear')
])
```

- Learning Rate Scheduler hinzugefügt, epochs = 50

```
# Compile
optimizer = Adam(learning_rate=0.001)
model.compile(optimizer=optimizer, loss='mse', metrics=['mae', 'mape'])

# Training with Learning Rate Scheduler
def scheduler(epoch, lr):
    return lr * 0.9
```

Optimierung des neuronalen Netzes

UNIVERSITÄT MANNHEIM
Fakultät für Betriebswirtschaftslehre

- Darstellung der Loss-Funktionen für Trainings- und Validierungsdatensatz

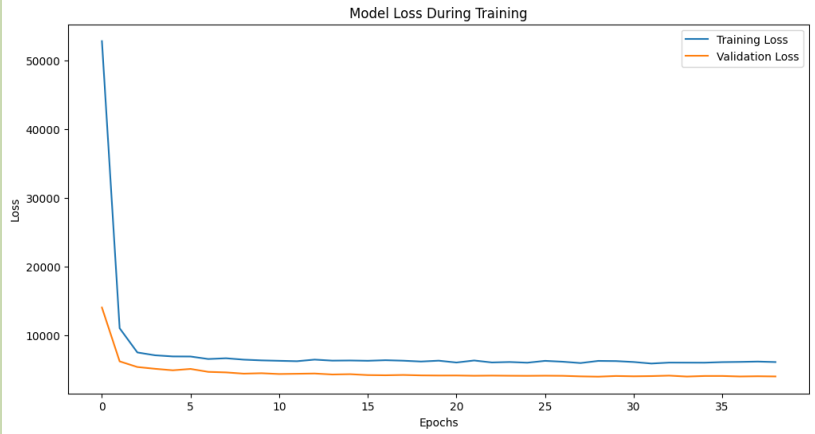
- MAPEs für den Trainings- und Validierungsdatensatz insgesamt

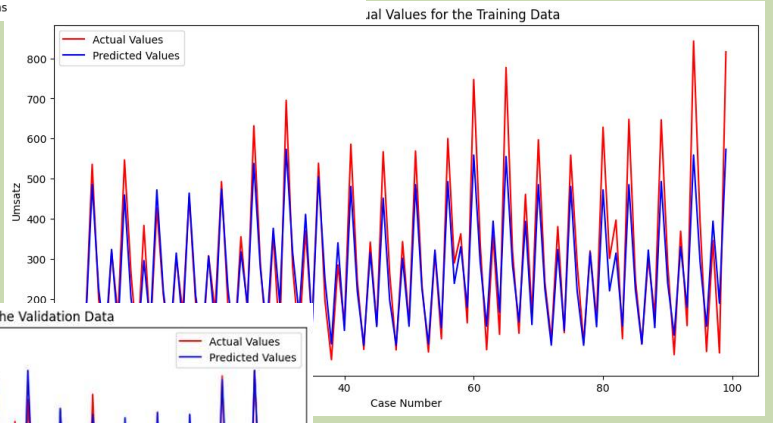
```
MAPE on the Training Data: 29.12%
MAPE on the Validation Data: 27.81%
```

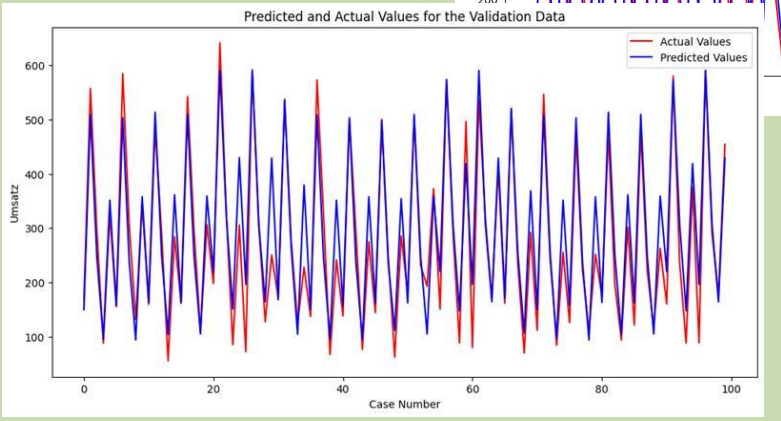
- MAPEs für jede Warengruppe einzeln

```
MAPE für Training Data:
Brot: 33.26%
Broetchen: 18.44%
Croissant: 22.73%
Konditorei: 29.04%
Kuchen: 16.26%
Saisonbrot: 189.84%
```

```
MAPE für Validation Data:
Brot: 29.75%
Broetchen: 22.96%
Croissant: 22.04%
Konditorei: 28.32%
Kuchen: 18.24%
Saisonbrot: 158.56%
```







„Worst Fail“

UNIVERSITÄT MANNHEIM
Fakultät für Betriebswirtschaftslehre

1

Lineares Modell: negativer R-squared

```
Mean Squared Error: 37799.071133370024
R-squared: -0.26031566281171625
Model Coefficients: [-9.08597778e+00 -1.59872116e-14 -6.33983762e+00 -4.36397897e+00
 -8.70572713e+00  5.12331823e-01 -4.71260964e+01 -1.30793096e+01
 -1.86484548e+01 -6.50800604e+01 -9.24441329e+00  1.29404963e+02
  2.37733715e+01  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00]
Intercept: 448.5990372583461
```

2

```
234/234 ━━━━━━━━━━━━━━━━━━━━ 0s 892us/step
56/56 ━━━━━━━━━━━━━━━━━━━━ 0s 911us/step
MAPE on the Training Data: 1112.91%
MAPE on the Validation Data: 651.84%
```

Neuronales Modell: sehr hoher MAPE

3

Neuronales Modell: Gerade als Vorhersage