Projekt - Telco Customer Churn

Analytische Informationssysteme
Dozent: Prof. Dr. Roland Müller

Serif Gören

Gliederung

01
Geschäftsverständnis
Geschäftsproblem & Zielsetzung

02
Datenverständnis
Beschreibung & Visualisierung der Daten, fehlende Daten

03
Datenvorbereitung
Auswahl & Ableitung von Merkmalen, Aufteilung in Trainings- und Testdaten

04
Modellbildung
Auswahl & Training des Modells, Anpassungen

05
Modellevaluierung
Bewertung & Vergleich der Modelle

06
Geschäftliche Nutzung
Modellanwendung zur Problemlösung

01

Geschäftsverständnis

Geschäftsproblem & Zielsetzung

Geschäftsproblem

- Datensatz von IBM
  → Fiktives Telekommunikationsunternehmen
  → 7043 Einträge mit 20 Features

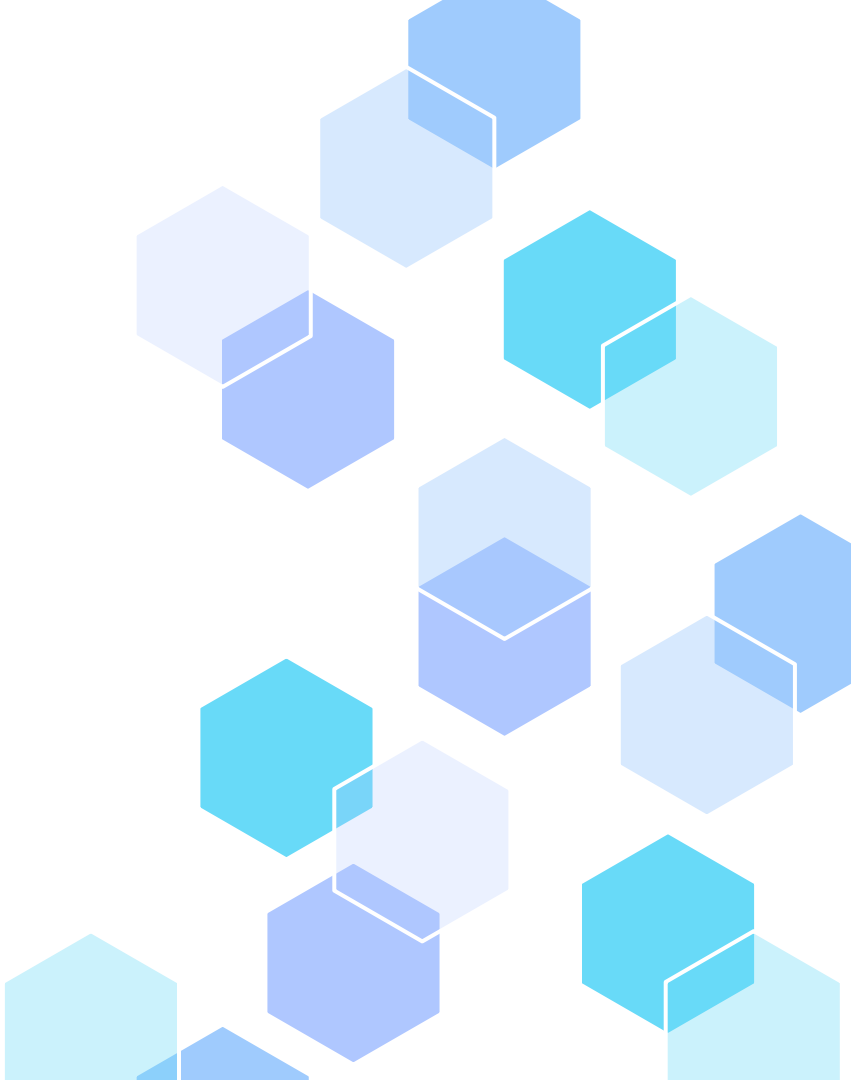
Zielsetzung:

- Kundenverhalten analysieren
- Abwanderungsfaktoren identifizieren
- Reduzierung der Abwanderung durch Vorhersagemodelle
- Entwicklung geeigneter Kundenbindungsprogramme zur präventiven Minimierung von Abwanderung
- Hauptfokus ist “Churn”-Spalte:
  → Hat der Kunde das Unternehmen in den letzten Monaten verlassen?

02

Datenverständnis

Beschreibung & Visualisierung der Daten, fehlende Daten



Features in Kategorien

Kategorien	Features (20)
Zielvariable	**Churn (Yes/ No)**
Demografische & Account Informationen	Gender, SeniorCitizen, Partner, Dependents
Dienstleistungen	PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies
Vertrag- und Zahlungsmethoden	Contract, PaperlessBilling, PaymentMethod, Tenure
Finanzielle Aspekte	MonthlyCharges, TotalCharges

Features & mögliche Werte

Demografische & Account Informationen (4):

- **Gender** (Male/ Female)
- **SeniorCitizen** (0=No/ 1=Yes)
- **Partner** (Yes/ No)
- **Dependents** (Yes/ No)

Dienstleistungen (9):

- **PhoneService** (Yes/ No)
- **MultipleLines** (Yes/ No/ No phone service)
- **InternetService** (DSL/ Fiber optic/ No)
- **OnlineSecurity** (Yes/ No/ No internet service)
- **OnlineBackup** (Yes/ No/ No internet service)
- **DeviceProtection** (Yes/ No/ No internet service)
- **TechSupport** (Yes/ No/ No internet service)
- **StreamingTV** (Yes/ No/ No internet service)
- **StreamingMovies** (Yes/ No/ No internet service)

Kategoriale Daten

Numerische Daten

Features & mögliche Werte

Vertrag- und Zahlungsmethoden (4):

- **Contract** (Month-to-month/ 1 year/ 2 year)
- **PaperlessBilling** (Yes/ No)
- **PaymentMethod** (Electronic check/ Mailed check/ Bank transfer (automatic)/ Credit card (automatic)
- **Tenure** (kontinuierlicher Wertebereich)

Finanzielle Aspekte (2):

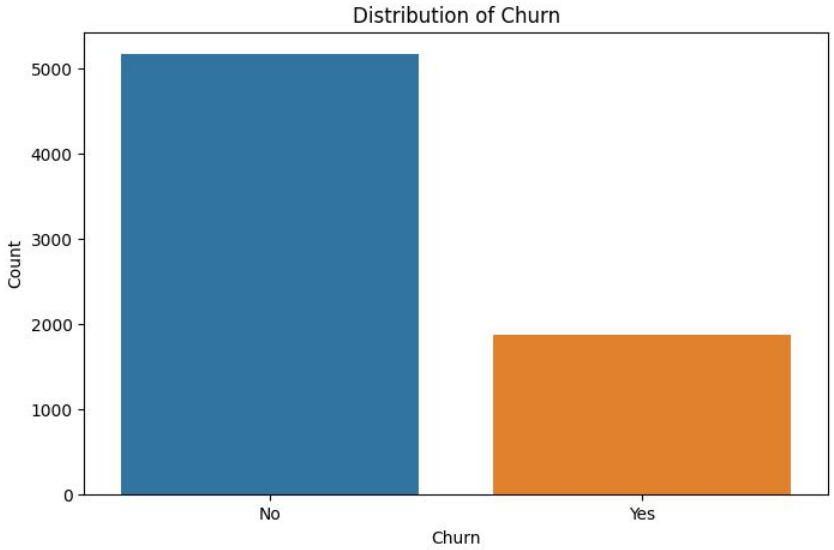
- **MonthlyCharges** (kontinuierlicher Wertebereich)
- **TotalCharges** (kontinuierlicher Wertebereich)

Kategoriale Daten
Numerische Daten

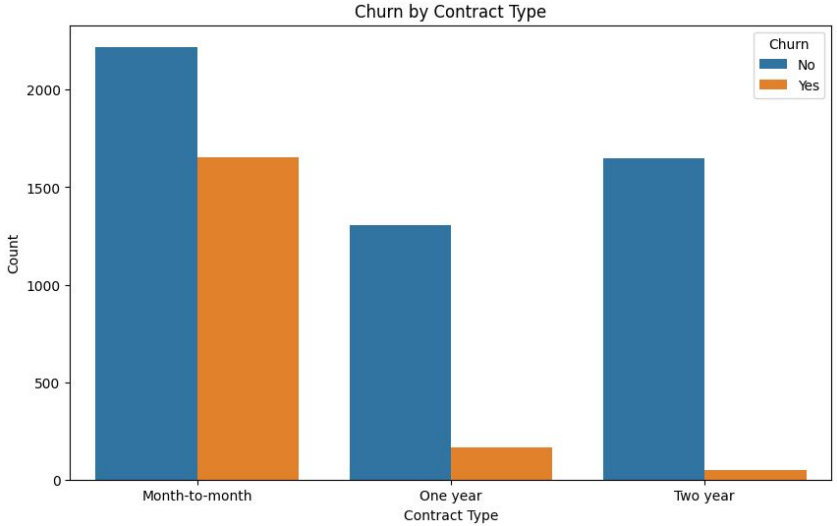
Datenvisualisierung

- **Balkendiagramme**
  → Ideal für kategoriale Daten
- **Histogramme**
  → Ideal für numerische Daten
- **Boxplots**
  → Nützlich für Ausreißer und Verteilung in numerischen Daten
- **Heatmaps**
  → Korrelation zwischen verschiedenen numerischen Merkmalen

Datenvisualisierung

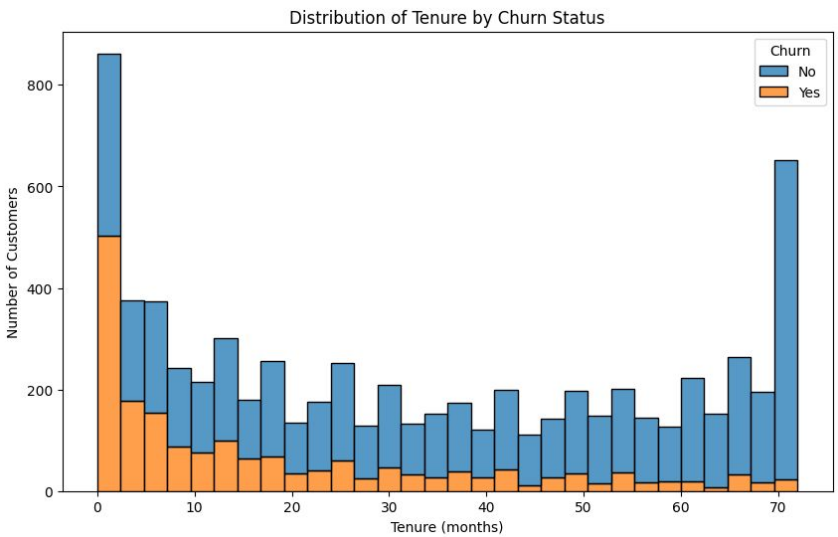


Kundenabwanderung

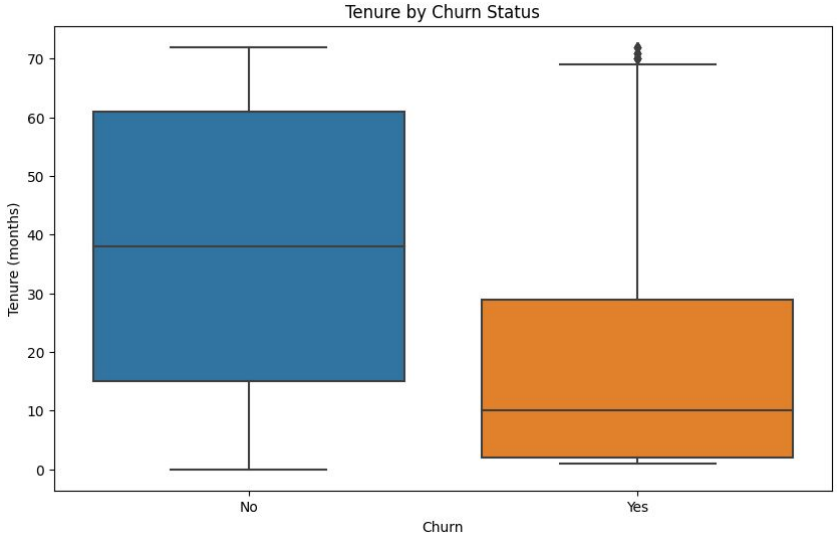


Vertragslaufzeit

Datenvisualisierung

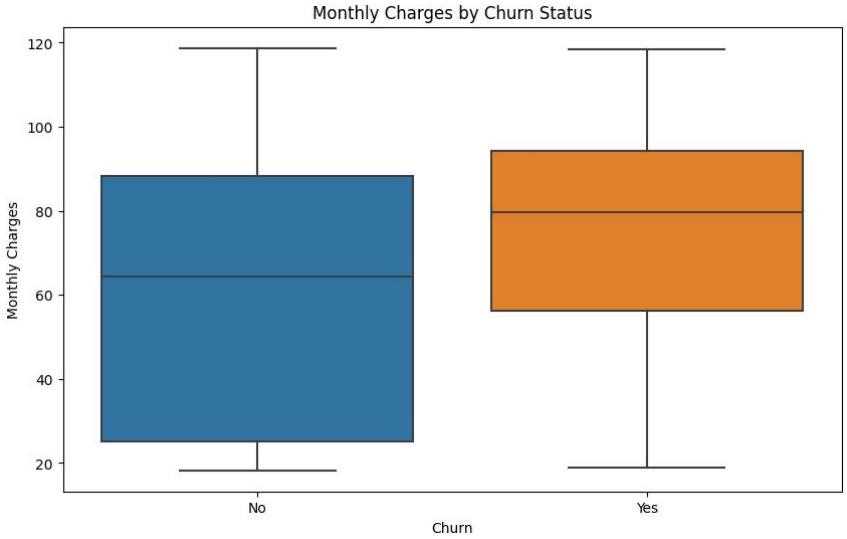


Vertragsdauer

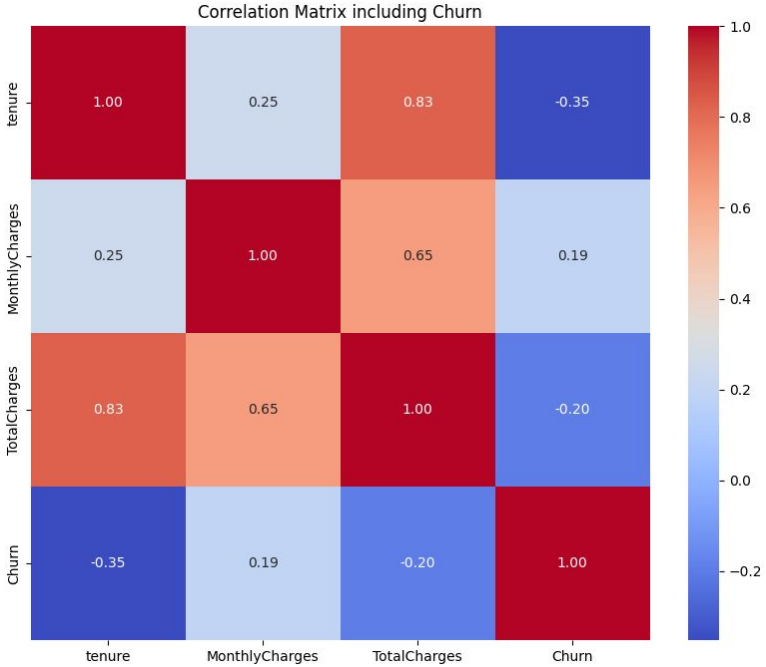


Vertragsdauer

Datenvisualisierung

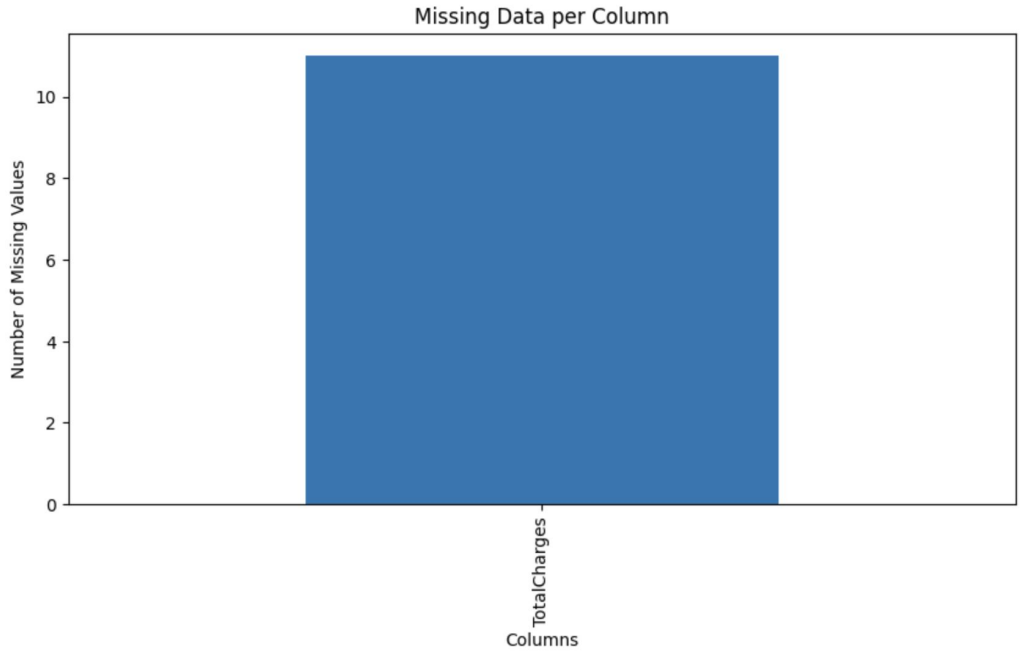


Kundenabwanderung



Heatmap mit
numerischen Daten

Fehlende Daten

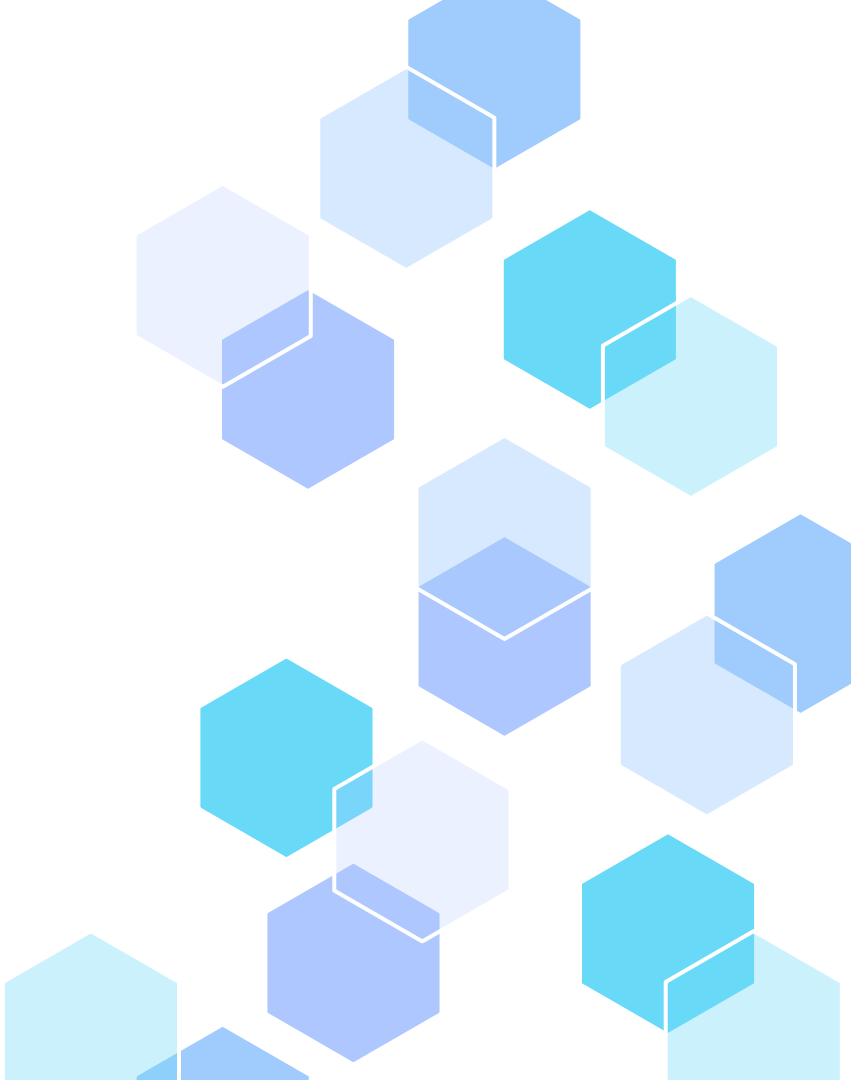


11 fehlende Werte bei
TotalCharges

03

Datenvorbereitung

Auswahl & Ableitung von Merkmalen,
Aufteilung in Trainings- und Testdaten



Auswahl von Features

Kriterien für die Auswahl von Features

- Features mit klarem Unterschied zwischen "Abgewandert" und "Nicht-Abgewandert" bei Balkendiagramm, Boxplot
- hohe Korrelation mit Zielvariablen "Abwanderung"
- Geschäftsverständnis

Auswahl von Features

Höhere Churn:

Partner	Depen-dents	Internet Service	Online Service	Con-tract	Pay-ment Method	Online Backup	Device Protec-tion	Tech Support	Paper-less Billing	Senior Citizen	Tenure	Monthly Char-ges	Total Char-ges
✗	✗	Glas-faser	✗	Monat-lich	elektr.	✗	✗	✗	✓	✓	kürzer	höher	geringer

Mittlere Churn:

Multiple Line
✓

Weniger relevant:

Gender	Phone Service	Stream-ing TV	Stream-ing Movies

Nicht relevant:

Cost-umer ID

Ableitung von Features

TotalExtraServices

Anzahl der abonnierten Zusatzdienste

MultipleServices

Mehrere Zusatzdienste (Binär)

CustomerEngagement Score

Anzahl abonnierten Zusatzdienste und Vertragsdauer (Score)

Aufteilung Trainings- und Testdaten

- **Erstellung der Pipeline**
  → Aufteilung in kategoriale & numerische Daten
- **ColumnTransformer** mit SimpleImputer, StandardScaler und OneHotEncoder
  → Kombination verschiedener Schritte zur Verarbeitung der Daten
- **Aufteilung der Daten:** 70/30

04

Modellbildung

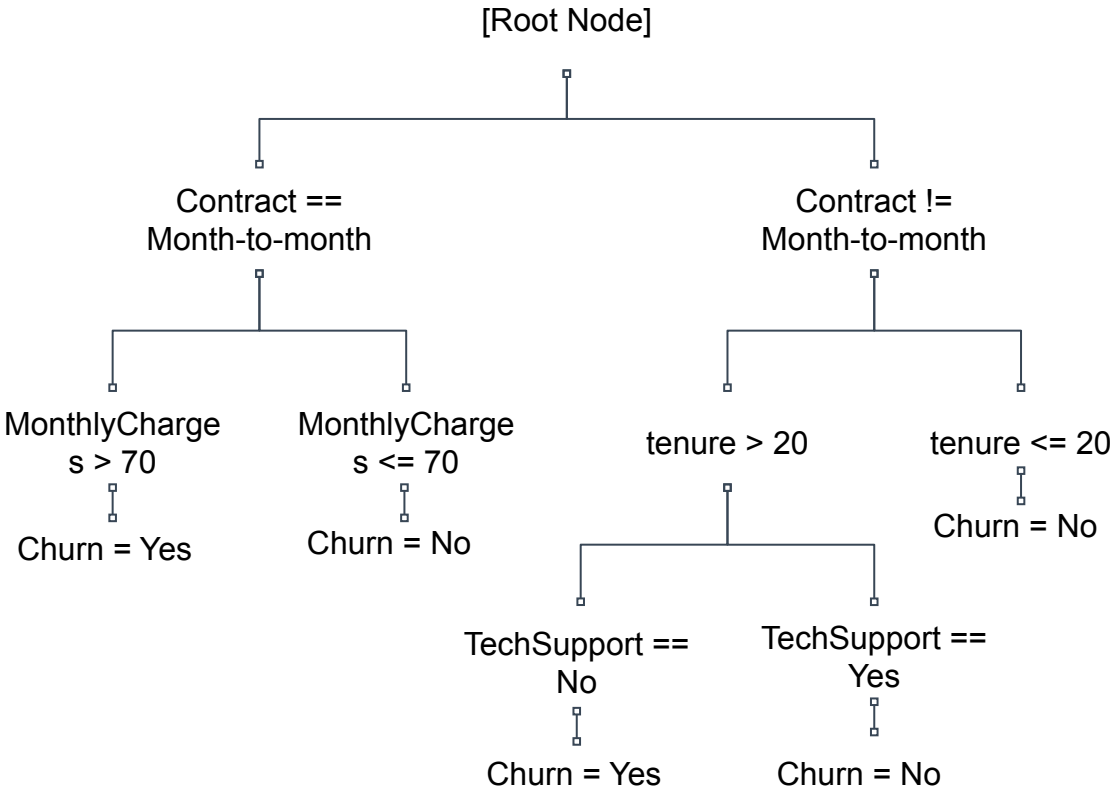
Auswahl & Training des Modells, Anpassungen

Entscheidungsbaum

- Vorhersagen werden basierend auf Datenmerkmalen erstellt
- Besteht aus Knoten und Verzweigungen
  - Knoten: Frage/Bedingung
  - Verzweigungen: Mögliche Antworten
- Endpunkte (Blätter) stellen die Entscheidung dar (Vorhersage)

Pro	Contra
Leicht verständlich	“Overfitting” von Trainingsdaten
Geringere Datenvorbereitung	Hohe Sensibilität
Verarbeitung numerischer und kategorialer Daten	Optimierung der Hyperparameter schwer möglich

Entscheidungsbaum



Random Forest

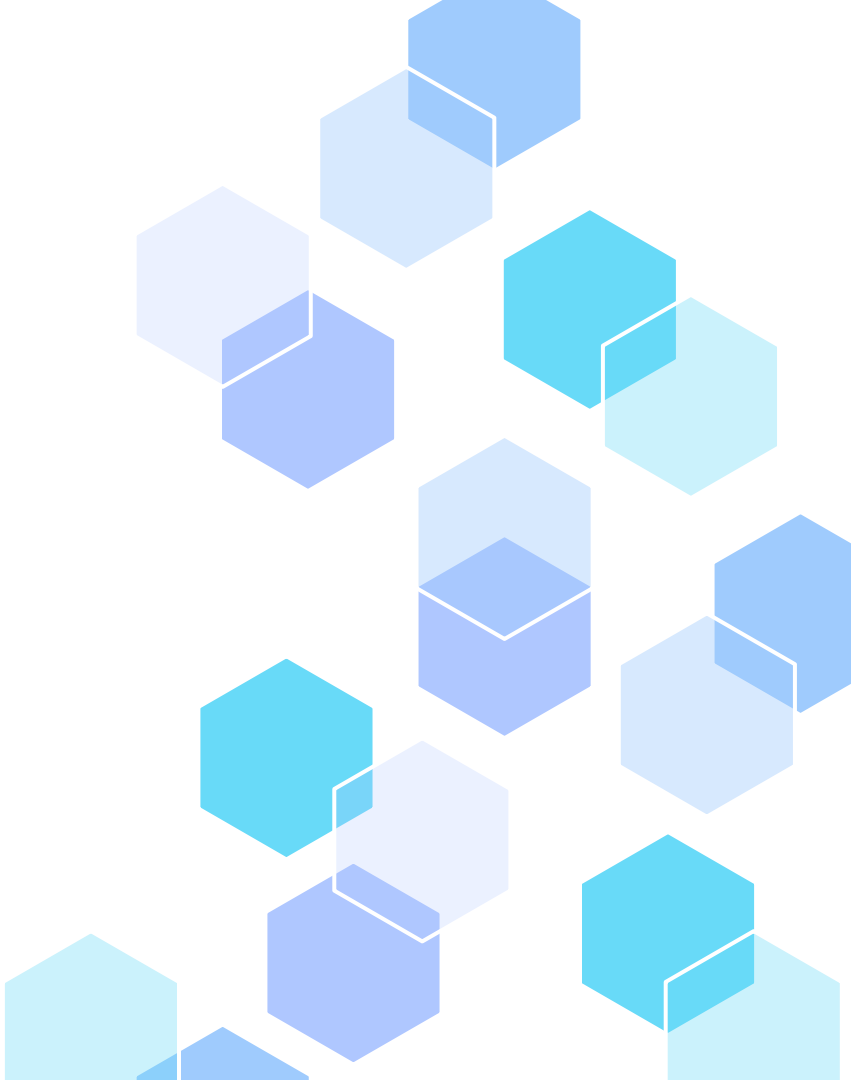
- Sammlung von vielen Entscheidungsbäumen
- Jeder Baum wird auf zufälligen Teilmenge der Daten trainiert
- Abstimmung von allen Bäumen bei Vorhersage
  → Häufigste Vorhersage wird ausgewählt

Pro	Contra
Hohe Genauigkeit	Schwerer zu interpretieren
Overfitting reduziert	Mehr Rechenleistung
Verarbeitung numerischer und kategorialer Daten	

05

Modellevaluierung

Bewertung & Vergleich der Modelle



Decision Tree – Evaluation

- **Genauigkeit:** 77,66%
- **Precision (No):** 83%
  → erkennt Kunden, die bleiben
- **Recall (Yes):** 50%
  → identifiziert nur Hälfte der abgewanderten Kunden
- **Precision (Yes):** 61%
  → 39% der Vorhersagen für Abwanderungen sind falsch

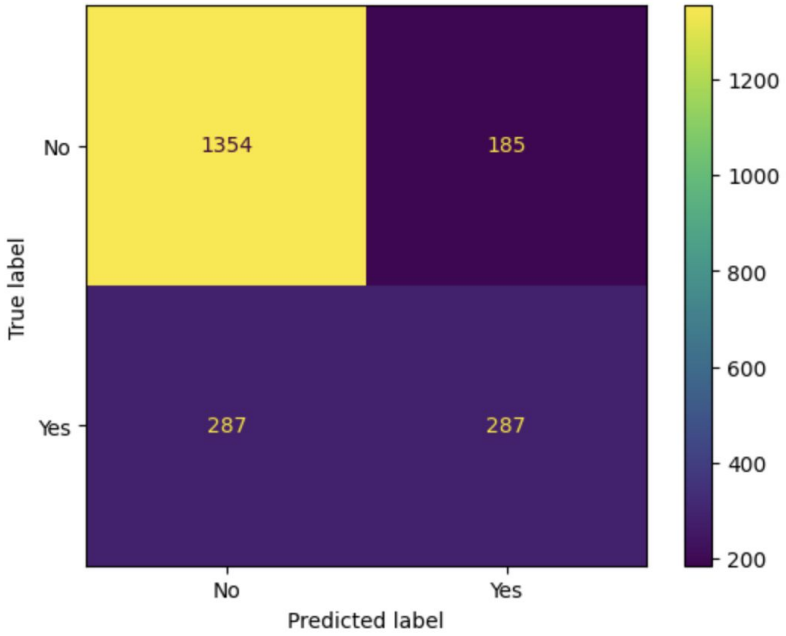
```
Accuracy of the best model on the test data set:
0.7766209181258874
Classification report of initial model:
              precision    recall  f1-score   support

          No       0.83      0.88      0.85      1539
         Yes       0.61      0.50      0.55       574

    accuracy                           0.78      2113
   macro avg       0.72      0.69      0.70      2113
weighted avg       0.77      0.78      0.77      2113
```

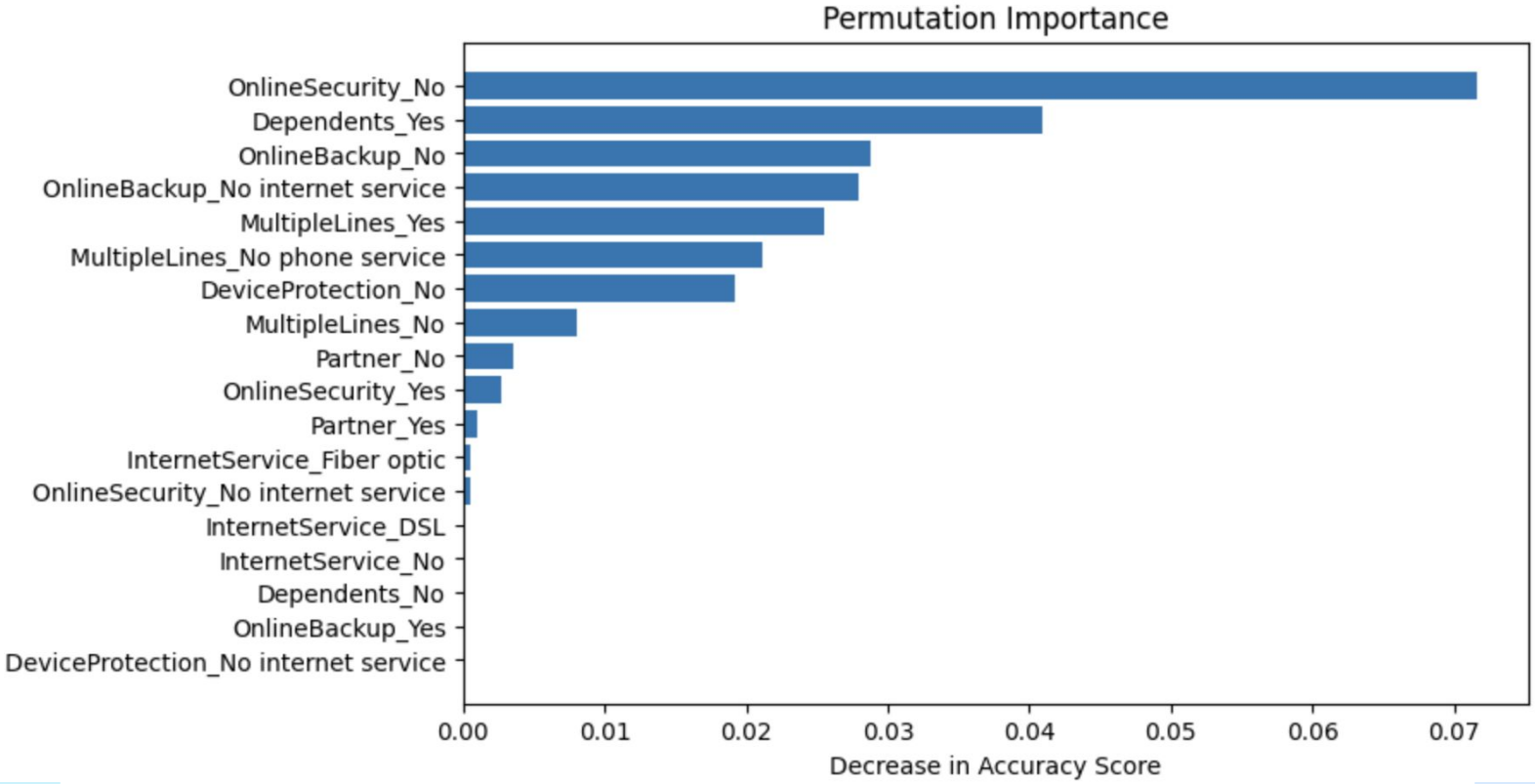
- Viele Abwanderungen werden nicht erkannt
- Ressourcen unnötig zur Kundenbindung verbraucht

Decision Tree – Konfusionsmatrix



Tatsächlich	Vorhergesagt: No	Vorhergesagt: Yes
No	TN	FP
Yes	FN	TP

Decision Tree – Permutation Importance



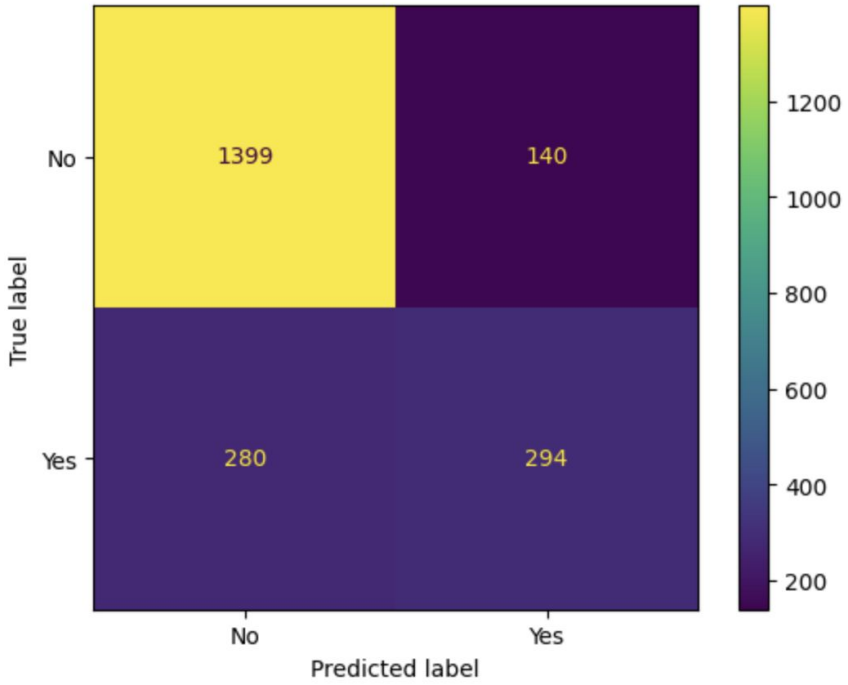
Random Forest – Evaluation

- **Genauigkeit:** 80,12%
- **Precision (Yes):** 68%
  → weniger Ressourcenverschwendung durch präzisere Vorhersagen für Abwanderung
- **Recall (Yes):** 51%
  → identifiziert etwas mehr als die Hälfte der abwandernden Kunden, aber viele Abwanderungen werden weiterhin nicht erkannt

Accuracy: 0.8012304779933743

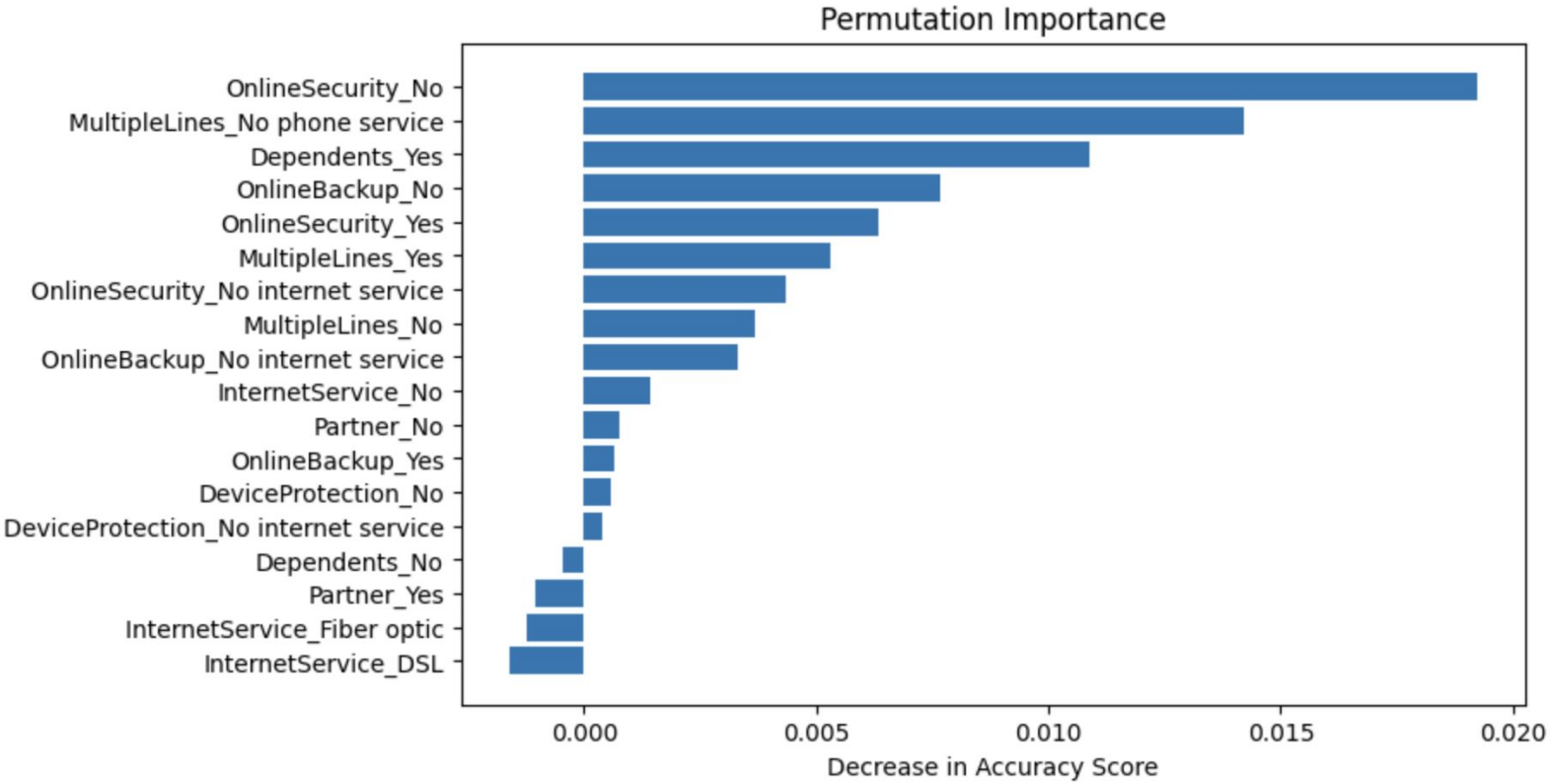
	precision	recall	f1-score	support
No	0.83	0.91	0.87	1539
Yes	0.68	0.51	0.58	574
accuracy			0.80	2113
macro avg	0.76	0.71	0.73	2113
weighted avg	0.79	0.80	0.79	2113

Random Forest – Konfusionsmatrix

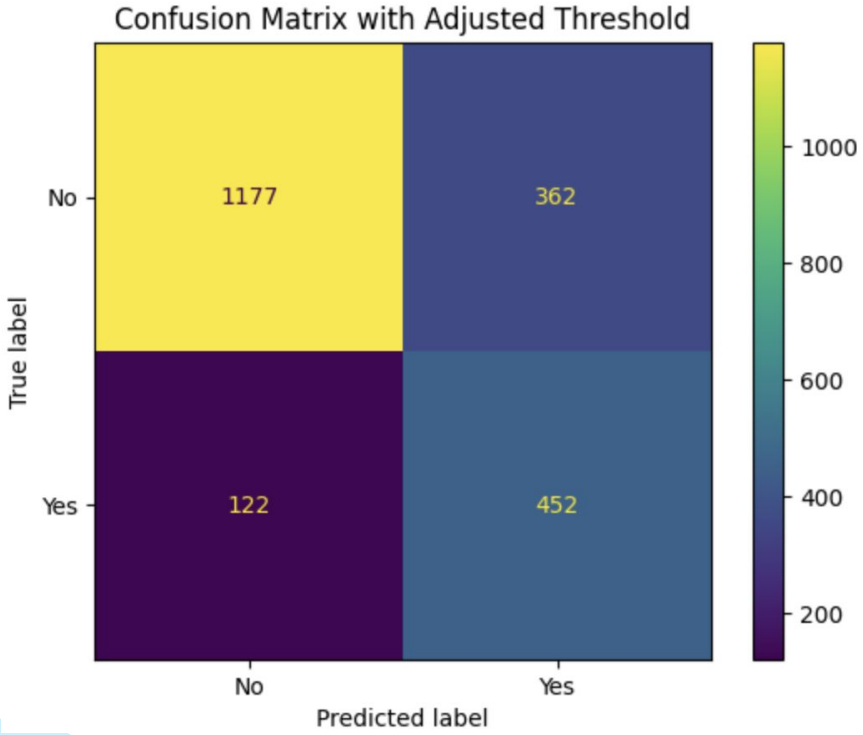


Tatsächlich	Vorhergesagt: No	Vorhergesagt: Yes
No	TN	FP
Yes	FN	TP

Random Forest - Permutation Importance



Random Forest – Schwellenwertanpassung



Accuracy with adjusted threshold: 0.7709417889256981

	precision	recall	f1-score	support
No	0.91	0.76	0.83	1539
Yes	0.56	0.79	0.65	574
accuracy			0.77	2113
macro avg	0.73	0.78	0.74	2113
weighted avg	0.81	0.77	0.78	2113

Vergleich der Modelle

- **Genauigkeit**: Random Forest (80,12%) zuverlässiger als Decision Tree (77,66%)
- **True Positives**: Random Forest erkennt mehr tatsächlich abwandernde Kunden
- **False Positives**: Random Forest hat weniger FP, was Ressourcen bei der Kundenbindung spart
- **Precision (Yes)**: Random Forest (68%) sorgt für weniger falsche Vorhersagen bei Abwanderung als Decision Tree (61%)
- **Recall (Yes)**: Beide Modelle haben einen ähnlichen Recall-Wert (51% vs. 50%)

→ Random Forest ist das leistungsstärkere Modell und besser geeignet für die Vorhersage von Kundenabwanderung

06

Geschäftliche Nutzung

Modellanwendung zur Problemlösung

Modellanwendung zur Problemlösung

- Beide Modelle helfen, abwandernde Kunden frühzeitig zu identifizieren
- Geeignete Maßnahmen können ergriffen werden, um die Kundenbindung zu stärken
- Reduzierung der falschen Vorhersagen spart Ressourcen
- Vorteil von Random Forest: Erkennt mehr richtige Kundenabwanderungen und macht weniger falsche Vorhersagen
- Gegenwirkung der Kundenabwanderung führt zu Umsatzsteigerung
- Frühzeitige Erkennung ermöglicht die Entwicklung geeigneter Modelle zur Kundenbindung und führt zu Wettbewerbsvorteil

Vielen Dank für eure Aufmerksamkeit!