Datathon

–Modul AI-

Team: Leon Bach, Nicole Maier, Michael Martin, Kevin Stutz, Felix Zentowski

Herman Hollerith Zentrum Böblingen
18. Juli 2023

Agenda

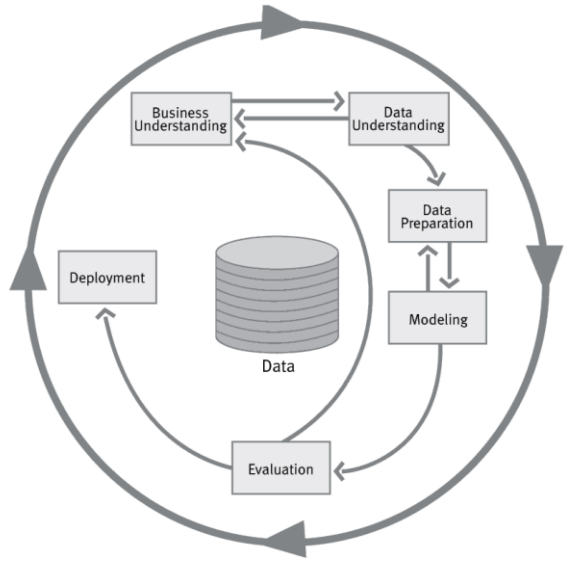
- Anforderungen
- Team-Work
- Business Understanding
- Data Understanding & Data Preparation
- Modeling
- Evaluation
- Deployment
- Services & Effekte

Anforderungen

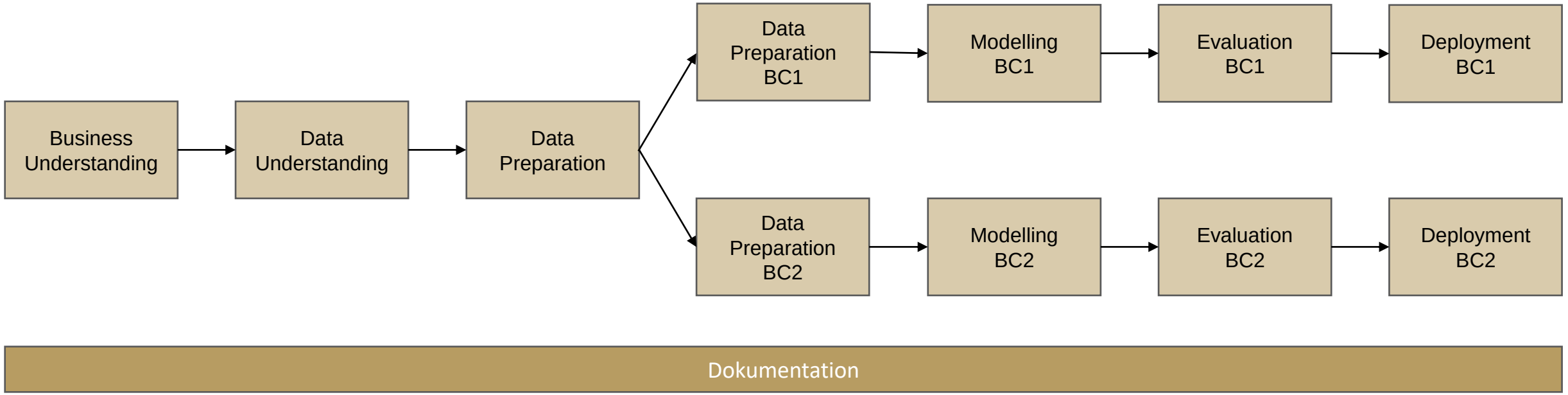
Ziel: Optimierung der HR-Strategie von Vertriebsmitarbeitenden: Steigerung des Engagement und der Loyalität

- Analyse relevanter Metriken
- Entwicklung neuer digitaler Services
- Empfehlungen basierend auf den Daten

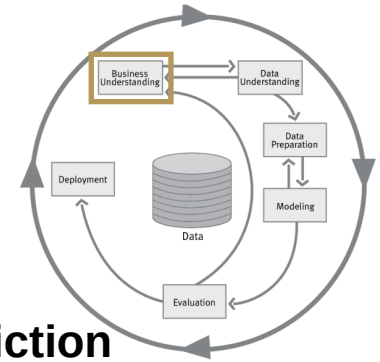
Team-Work



- Das Vorgehen orientiert sich am CRISP-DM Cycle.
- Nach gemeinsamer Data Preparation erfolgt die Aufteilung in die unterschiedlichen Business Cases (BC).
- Die Dokumentation erfolgt kontinuierlich während des gesamten Prozesses.



Business Understanding



Business Case 1: Quit Prediction

Problem:

Fluktuation von Leistungstragenden

Ziel:

Fluktuation minimieren durch Identifikation möglicher Maßnahmen zur Mitarbeiterbindung

Anforderung:

Zusammenhang zwischen Fluktuation und Merkmalen herstellen

Kriterium: Recall

Business Case 2: Basic Salery Prediction

Problem:

Bestimmung des angemessenen Einstiegsgehalts

Ziel:

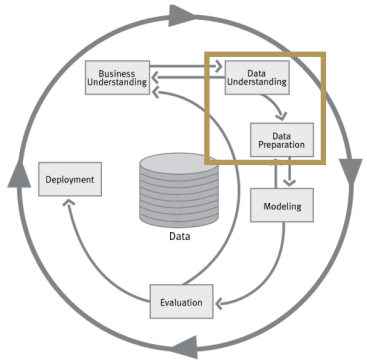
Datenbasierende Bestimmung des Einstiegsgehalts in Abhängigkeit bestehender Mitarbeitender

Anforderung:

Zusammenhang zwischen Gehaltshöhe und anderen Merkmalen herstellen

Kriterien: R^2, prozentualer Fehler, (MSE)

Data Understanding & Data Preparation Überblick

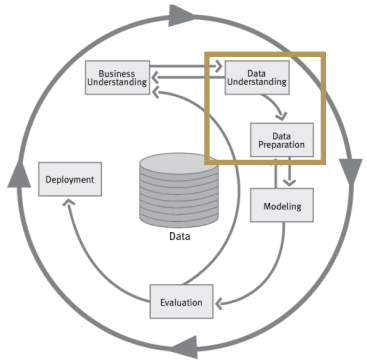


Anzahl der Datensätze: 1.250 mit 21 Features

<u>Deskriptive Analyse:</u>

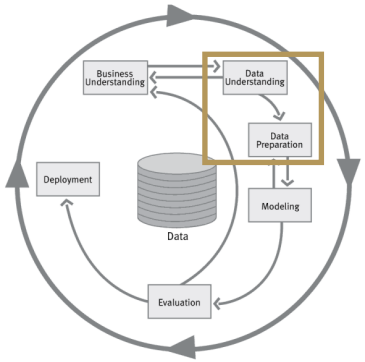
	Age	Gender	City	Position	Quit	YearsOverall	YearsGer	Seniority	MainTech	OtherTech	BasicSalary	SalaryBonus	LY_Salary	LY_Bonus	VacDays	EmplStatus	Contract	Language	CusSize	CusArea
count	1226.000000	1243	1253	1247	1234	1237	1221	1241	1126	1096	1.252000e+03	828	8.840000e+02	613	1184	1235	1223	1236	1234	1227
unique	NaN	3	112	148	2	48	53	24	251	562	NaN	167	NaN	130	45	11	3	14	5	63
top	NaN	Male	Berlin	Software Engineer	No	10	2	Senior	Python	Javascript / Typescript	NaN	0	NaN	0	30	Full-time employee	Unlimited contract	English	1000+	Product
freq	NaN	1049	681	387	1005	138	195	565	214	44	NaN	227	NaN	200	488	1189	1158	1020	448	759
mean	32.509788	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.034313e+07	NaN	6.329228e+05	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	5.663804	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.826189e+09	NaN	1.681458e+07	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	20.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000100e+04	NaN	1.100000e+04	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	29.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5.895000e+04	NaN	5.500000e+04	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	32.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	7.000000e+04	NaN	6.500000e+04	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	35.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.000000e+04	NaN	7.500000e+04	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	69.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000e+11	NaN	5.000000e+08	NaN	NaN	NaN	NaN	NaN	NaN	NaN

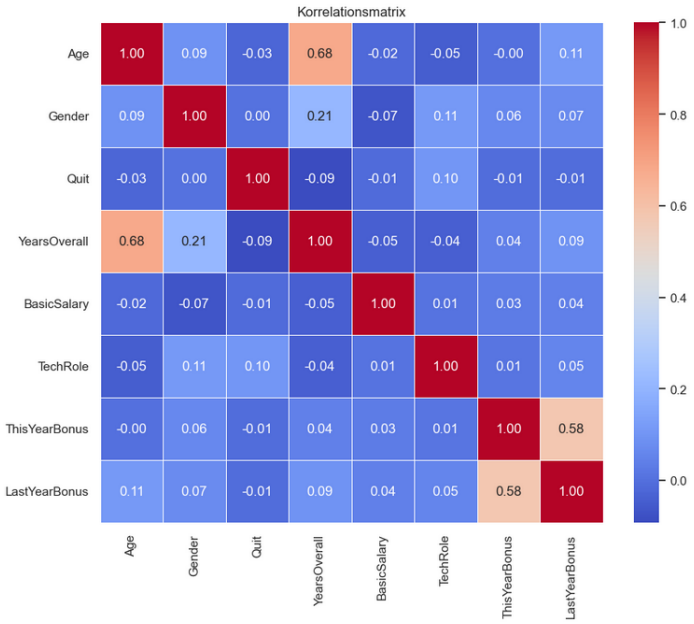
Data Understanding & Data Preparation
Konsistenz und Vollständigkeit der Daten



- Fehlende Werte im Datensatz enthalten z.B. Age, Gender
  - Medianimputation bei Age, Löschen der Werte bei Gender
- Identifikation falscher Werte z.B. YearsOverall
  - Löschen/Ersetzen der falschen Werte
- Merkmale mit einer großen Anzahl an Kategorien
  - Verkleinerung durch Feature Engineering z.B. TechRole, Seniority
  - Löschen von Spalten z.B. City, VacDays
- Große Spannweite bei nummerischen Features
  - Ausreißerbereinigung z.B. YearsOverall
- Umbennung der Spalte “Position” aufgrund Leerzeichen
- Vereinheitlichung der Datentypen Float und Integer

Data Understanding & Data Preparation
Konsistenz und Vollständigkeit der Daten

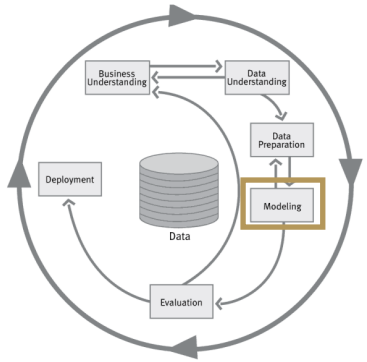




Korrelationsmatrix: Bereinigung erfolgt modellspezifisch.

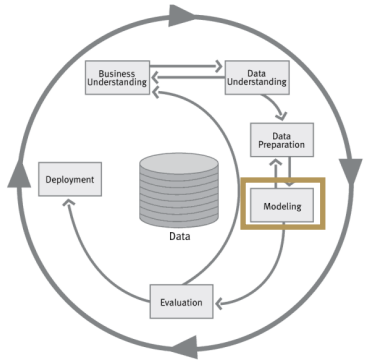
Ergebnis: 3 kategoriale Features, 8 nummerische Features

Modeling
Business Case 1

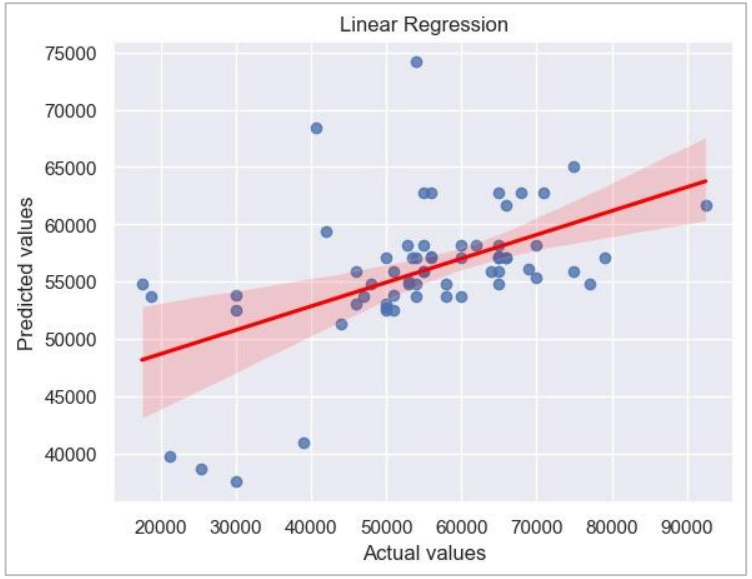


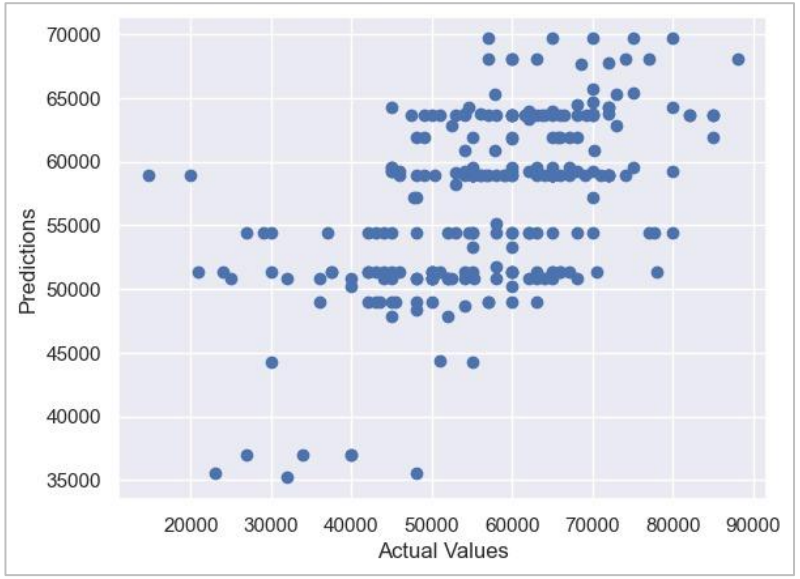
- Dummy Variablen erstellt
- Oversampling durchgeführt nach train test split
- Train/Test Split: 80/20
- Getestete Verfahren: Logistische Regression, Decision Tree, Random Forest, XG Boost
- Overfitting
  - Insbesondere bei den Baumbasierten Verfahren starkes Overfitting -->
  - Daraufhin wurden hyperparameter verändert. z.B. Baumtiefe angepasst
  - Bestes erreichtes Modell war RandomForest mit einer Baumtiefe von 5

Modeling
Business Case 2



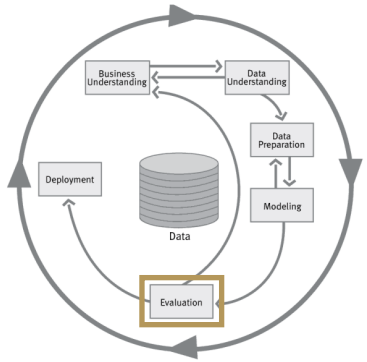
- Getestete Verfahren: Lineare Regression, Random Forest Regressor
- Train/Test Split: 80/20
- Spezifische Datenbreinigung:
  - Gender, CusSize, ThisYearBonus, LastYearBonus entfernt
  - Quit = 0, Level = Employee
- Multikollinearität:
  - VIF > 10: Age
- Dummy-Variablen
- Feature Scaling:
  - Standardisierung
- Hyperparameter Optimierung:
  - Keine Verbesserung





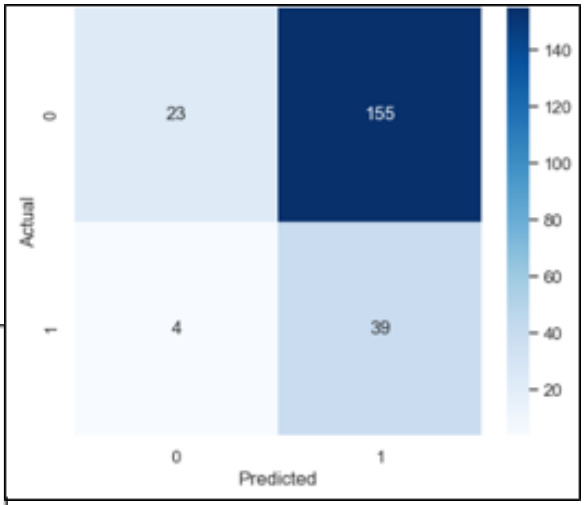
DBM
DIGITAL BUSINESS MANAGEMENT

Evaluation
Business Case 1



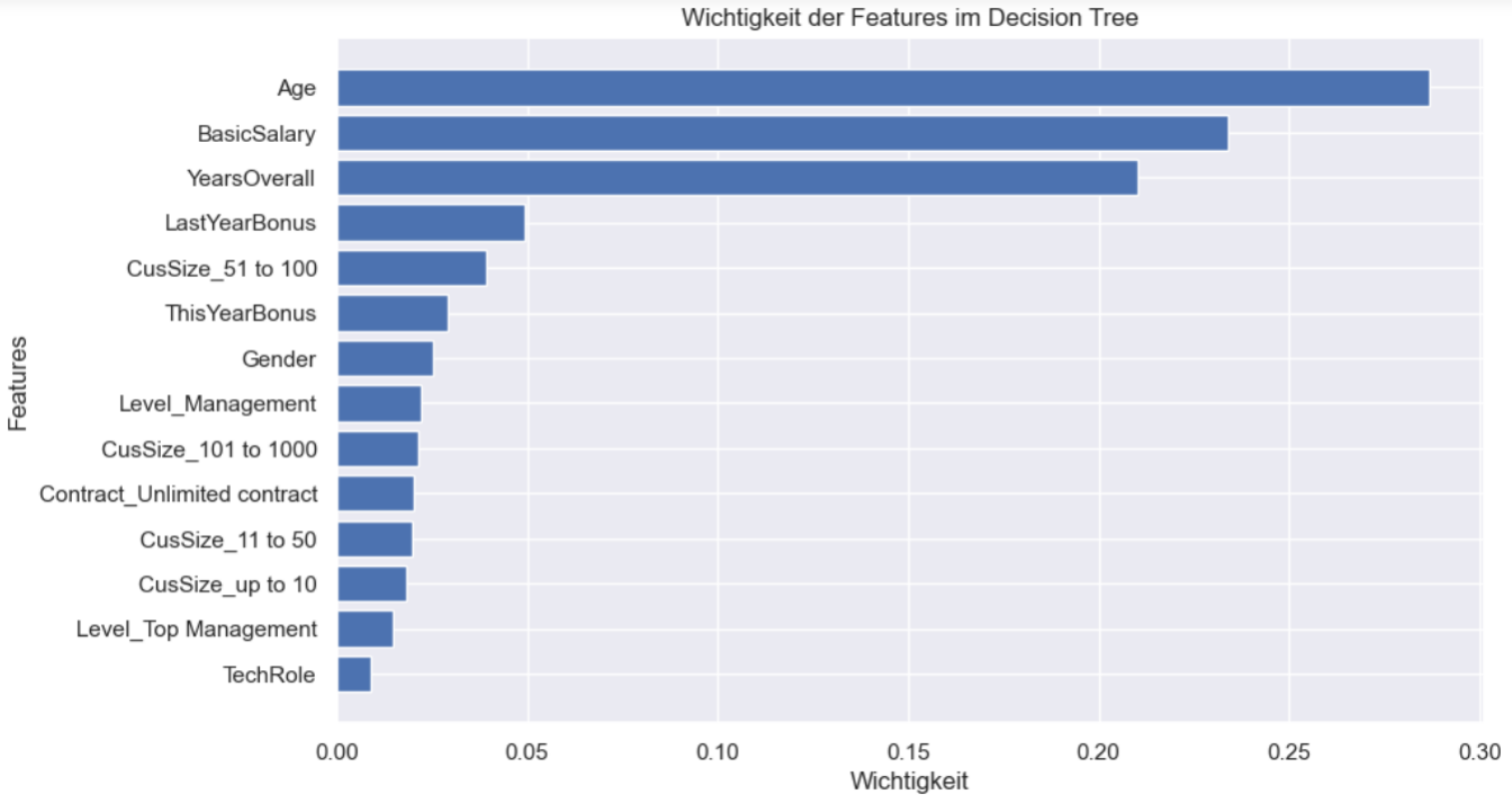
Modell	Log.Regress.	DesicionTree	RandomForest	XG Boost
F1-score(class1)	35%	35%	39%	36%

- Ergebnis
  - Random Forest mit Baumtiefe von 5 ist bestes Modell mit f1 score von 39%
  - Recall für diesen BC entscheidender Parameter.
  - Durch eine Übergewichtung von 2:1 auf der Klasse "1" --> Recall von 91%
  - Precision sinkt dadurch auf 20%
    - Es gibt 155 falsch positiv klassifizierte Personen
    - Aber 39 abwandernde werden erkannt



	precision	recall	f1-score	support
0	0.85	0.13	0.22	178
1	0.20	0.91	0.33	43
accuracy			0.28	221
macro avg	0.53	0.52	0.28	221
weighted avg	0.73	0.28	0.24	221

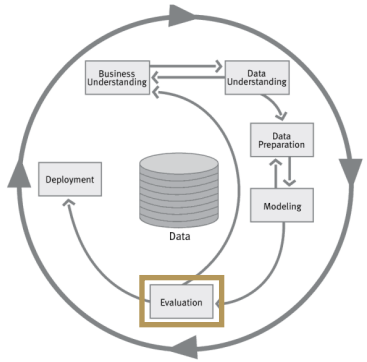
Business Case 1 – Evaluation



Gründe für Accuracy:

- Sehr hohe ungleichverteilung der Zielvariable.
- Dem kann durch scaling entgegengewirkt werden aber dennoch gibt es deutlich weniger Fälle, in welchen Mitarbeitende abwandern

Evaluation
Business Case 2

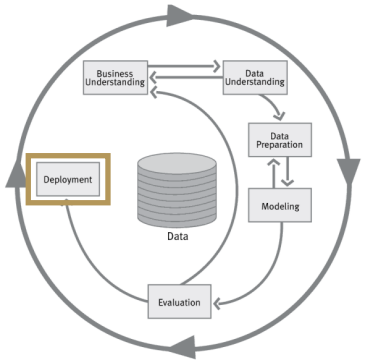


- Ergebnis:
  - Lineare Regression: R^2 = 0,24, prozentualer Fehler = 16,7%
  - Random Forest Regressor: R^2 = 0,34, prozentualer Fehler = 16,6% → bestes Modell
- Feature Importance:

	Coeffecient
Quit	0.000000
YearsOverall	0.770763
TechRole	0.062429
Contract_Unlimited contract	0.166808

- Gründe für hohen prozentualen Fehler:
  - Stark schwankende Werte des Gehalts
  - Zu geringe Anzahl verfügbarer Daten
- Optimierungsmöglichkeiten:
  - Tests weiterer Modelle (Support Vector Regression, Neural Network)
  - Unterteilen der Gehälter in Kategorien

Deployment
Business Case 1



- Bereitstellung an die Personalabteilung, um Maßnahmen zur Mitarbeiterbindung einzusetzen, bevor der Mitarbeitende abwandert

<u>Integration in das bestehende Personalmanagementsystem:</u>

1. Extraktion des Modells
2. Integration auf Flaskserver
3. API-Abfrage über POST
4. Anzeigen im Personalsystem

<u>Limitationen und Annahmen:</u>

Language wird aus ethischen Gründen nicht berücksichtigt

Deployment
Business Case 2

- Bereitstellung an die Personalabteilung für die Bestimmung des Einstiegsgehalts

Integration in die bestehende Infrastruktur:

1. Extraktion des Modells
2. Integration auf Flaskserver
3. Erstellen eines Formulars im Intranet
4. Senden der Daten an Server
5. Rückgabe des vorgesagten Gehalts

<u>Limitationen und Annahmen:</u>

- Gender wird aus ethischen Gründen nicht berücksichtigt
- Nur für Employees

Services & Effekte

	Business Case 1	Business Case 2
Services	Personalentwicklungspläne, Predictive Analytics	Datenanalyse-Dashboard, Einstiegsgehaltsempfehlungen
Effekte	Erhöhte Mitarbeiterbindung, Kosteneinsparungen	Verbesserte Mitarbeiterbindung, Datenbasierte Entscheidungsfindung

Business Value:

Verbesserte Entscheidungsfindung

Potentielle Limitationen:

- Datenabhängigkeit
- Subjektivität in Daten aufgrund menschlicher Entscheidungen in Daten
- Ethische Überlegungen (Diskriminierung: Geschlecht, Muttersprache; Datenschutz)