Explainable AI

LIME & RBE auf Random Forest
mit Adult Census Datensatz

11.4.2025

Datensatz

Analyse der Daten

Adult Census Datensatz

- Enthält demografische Informationen aus der US-Volkszählung
- Soll Einkommen einer Person vorhersagen (<|>50000$ pro Jahr)

Datensatzauswahl

- Gut strukturiert und saubere Daten
- Weit verbreiteter Benchmark in der Forschung und im maschinellen Lernen –> wird in vielen Studien verwendet
- Vielfalt der Variablen Typen –> enthält sowohl kategorische als auch numerische Variablen
- Realistisches Datenset –> gut geeignet für praxisnahe Anwendungen

Enthaltene Daten im Datensatz

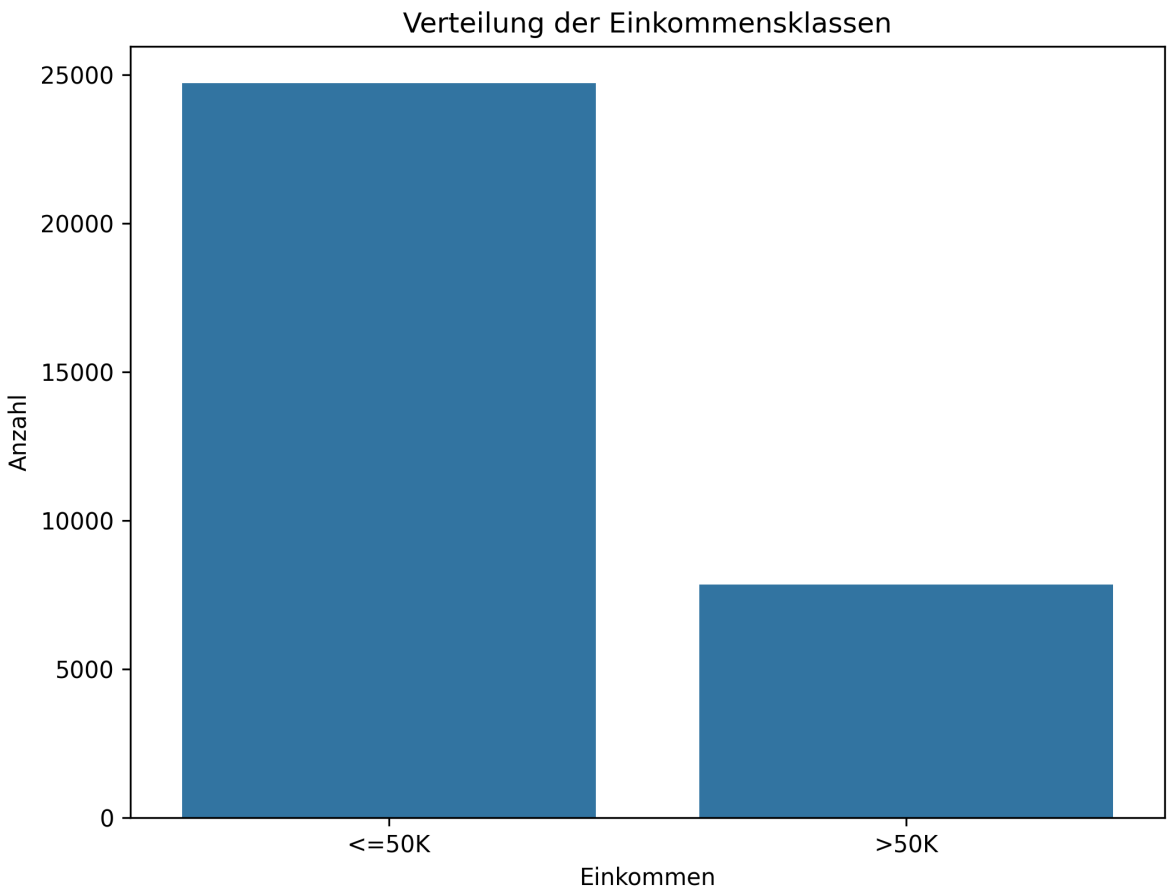
Column Name	Data Type
age	int64
workclass	object
fnlwgt	int64
education	object
education.num	int64
marital.status	object
occupation	object
relationship	object
race	object
sex	object
capital.gain	int64
capital.loss	int64
hours.per.week	int64
native.country	object
income	object

Analyse des Datensatzes

- Datensatzgröße
- Anzahl und Art der Datentypen
- Fehlende Werte in den einzelnen Spalten
- Statistische Zusammenfasung

	age	fnlwgt	education.num	capital.gain	capital.loss	hours.per.week
count	32561.000000	3.256100e+04	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	1.897784e+05	10.080679	1077.648844	87.303830	40.437456
std	13.640433	1.055500e+05	2.572720	7385.292085	402.960219	12.347429
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178270e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783560e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370510e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

Prozentuale Verteilung der Personen im Datensatz



Datenvorverarbeitung

Auffüllen von fehlenden Werten, Umwandlung von Variablen

Überprüfung auf fehlende oder ungültige Werte

```python
# Überprüfen auf fehlende Werte oder '?'
for col in df.columns:
    missing_count = df[df[col] == '?'].shape[0]
    if missing_count > 0:
        print(f"Spalte '{col}' hat {missing_count} Einträge mit '?'")
```

```
Spalte 'workclass' hat 1836 Einträge mit '?'
Spalte 'occupation' hat 1843 Einträge mit '?'
Spalte 'native.country' hat 583 Einträge mit '?'
```

Ersetzung von fehlenden/ ungültigen Werten (vorher)

	age	workclass	fnlwgt	education	education.num	marital.status	occupation
0	90	?	77053	HS-grad	9	Widowed	?
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial
2	66	?	186061	Some-college	10	Widowed	?
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct
4	41	Private	264663	Some-college	10	Separated	Prof-specialty

Ersetzung von fehlenden/ ungültigen Werten (nachher)

	age	workclass	fnlwgt	education	education.num	marital.status	occupation
0	90	Private	77053	HS-grad	9	Widowed	Prof-specialty
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial
2	66	Private	186061	Some-college	10	Widowed	Prof-specialty
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct
4	41	Private	264663	Some-college	10	Separated	Prof-specialty

Umwandlung der Variablen

- Von Katergorischen Variablen zu Numerischen
- Beispiel Geschlecht:
  - Mann -> 1
  - Frau -> 0
- Wird benötigt um mit dem Datensatz weiter zu arbeiten zu können -> ML-Modelle benötigen numerische Werte um sinnvolle Berechnungen durchführen zu können

ML-Modell

Auswahl, Beschreibung, Hyperparameter, Güte

Auswahl eines geeigneten Modells

Kriterium	**Unsere Situation**
Art der Daten	Strukturiert
Datenmenge	~32.500 Einträge
Komplexität	13-dimensionale Daten –> eher komplex
Ziel und Art der Klassifizierung	Binäre Klassifizierung (>/< 50.000$ Einkommen)

–> Gute Voraussetzungen für **Random Forest**

Random Forest

- **Ensemble-Learning-Verfahren**, bestehend aus vielen Entscheidungsbaummodellen (bei uns 100 Bäume)
- **Training:** Jeder Baum wird auf zufälliger Teilmenge der Daten und zufälliger Teilmenge der Features trainiert
- **Vorhersagen:** Ergebnisse aller Bäume werden aggregiert, um eine finale Vorhersage zu erhalten
- Keine Aktivierungsfunktionen oder Layer

Random Forest

Hyperparameter

> *Parameter, die vor dem Training festgelegt werden müssen*

Hyperparameter	Bedeutung
n_estimators	Anzahl an Bäumen
max_depth	Maximale Tiefe eines Baumes
min_samples_split	Mindestanzahl an Samples, um ein Split durchzuführen
min_samples_leaf	Mindestanzahl an Samples in einem Blatt

Hyperparameter-Tuning

> *Optimierung der Hyperparameter, um die Modellgüte zu verbessern*

```python
# Hyperparameter-Grid definieren
param_grid = {
    'n_estimators': [50, 100],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}

# GridSearchCV
grid_search = GridSearchCV(RandomForestClassifier(random_state=42), param_grid, cv=3, scoring='acc
grid_search.fit(X_train, y_train)

# Beste Parameter
print("Beste Parameter:")
print(grid_search.best_params_)
```

```
{
    'max_depth': None,
    'min_samples_leaf': 2,
    'min_samples_split': 2,
    'n_estimators': 100
}
```

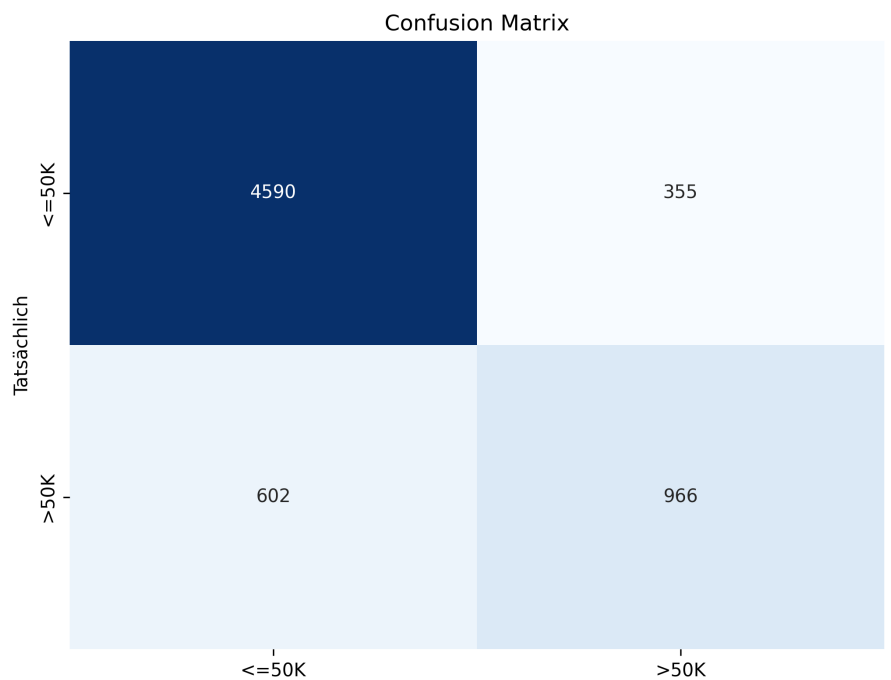
Hyperparameter-Tuning

```
# Hyperparameter-Grid definieren
param_grid = {
    'n_estimators': [50, 100],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}

# GridSearchCV
grid_search = GridSearchCV(RandomForestClassifier(random_state=42), param_grid, cv=3, scoring='acc
grid_search.fit(X_train, y_train)

# Beste Parameter
print("Beste Parameter:")
print(grid_search.best_params_)

# Bestes Modell
best_rf_model = grid_search.best_estimator_

# Vorhersagen mit dem besten Modell
y_pred_best = best_rf_model.predict(X_test)

# Modellleistung evaluieren
```

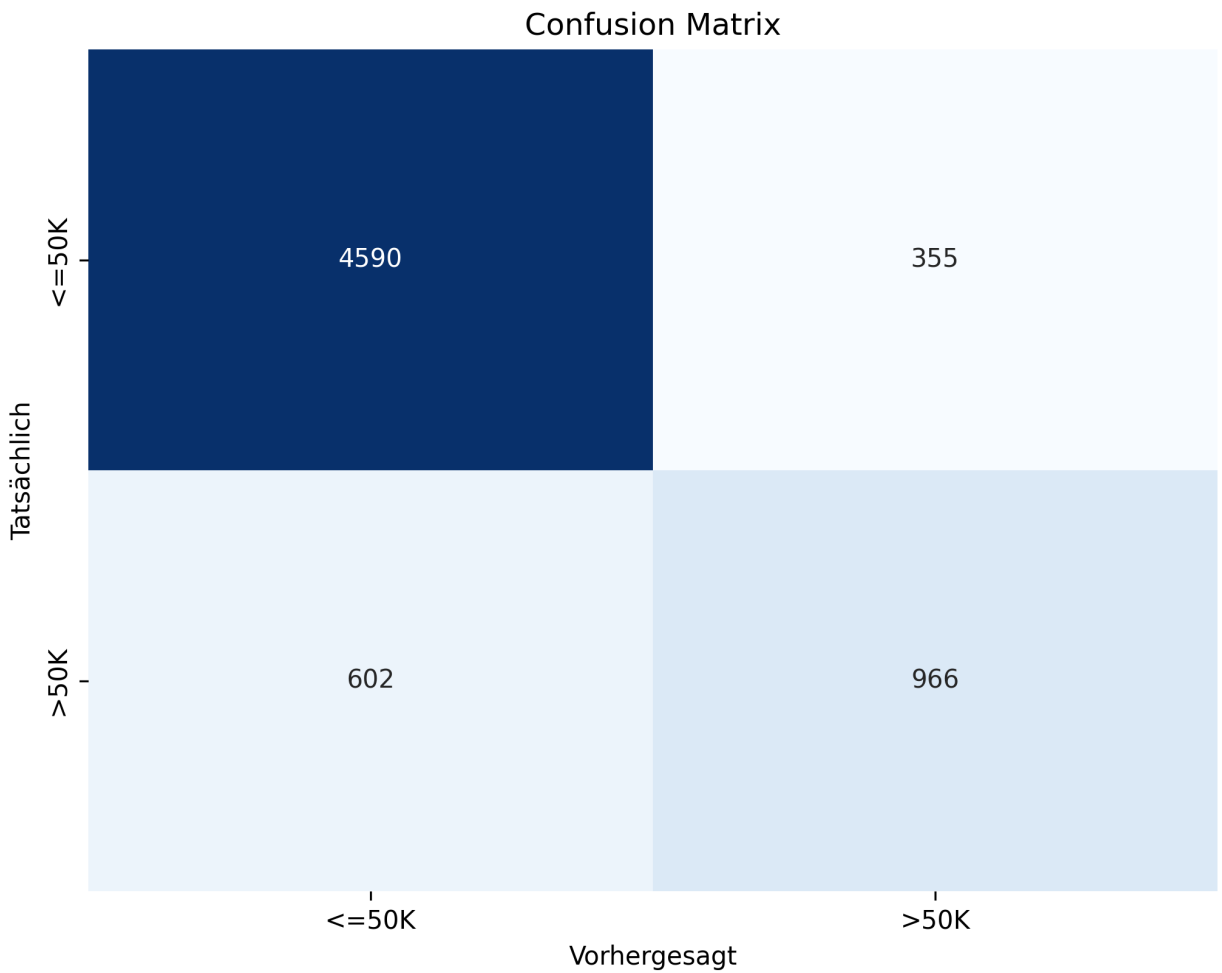
Modellgüte

```
{
  'Accuracy': 0.8607,
  'Precision': 0.7613,
  'Recall': 0.6142,
  'F1 Score': 0.6798
}
```

Konfusionsmatrix

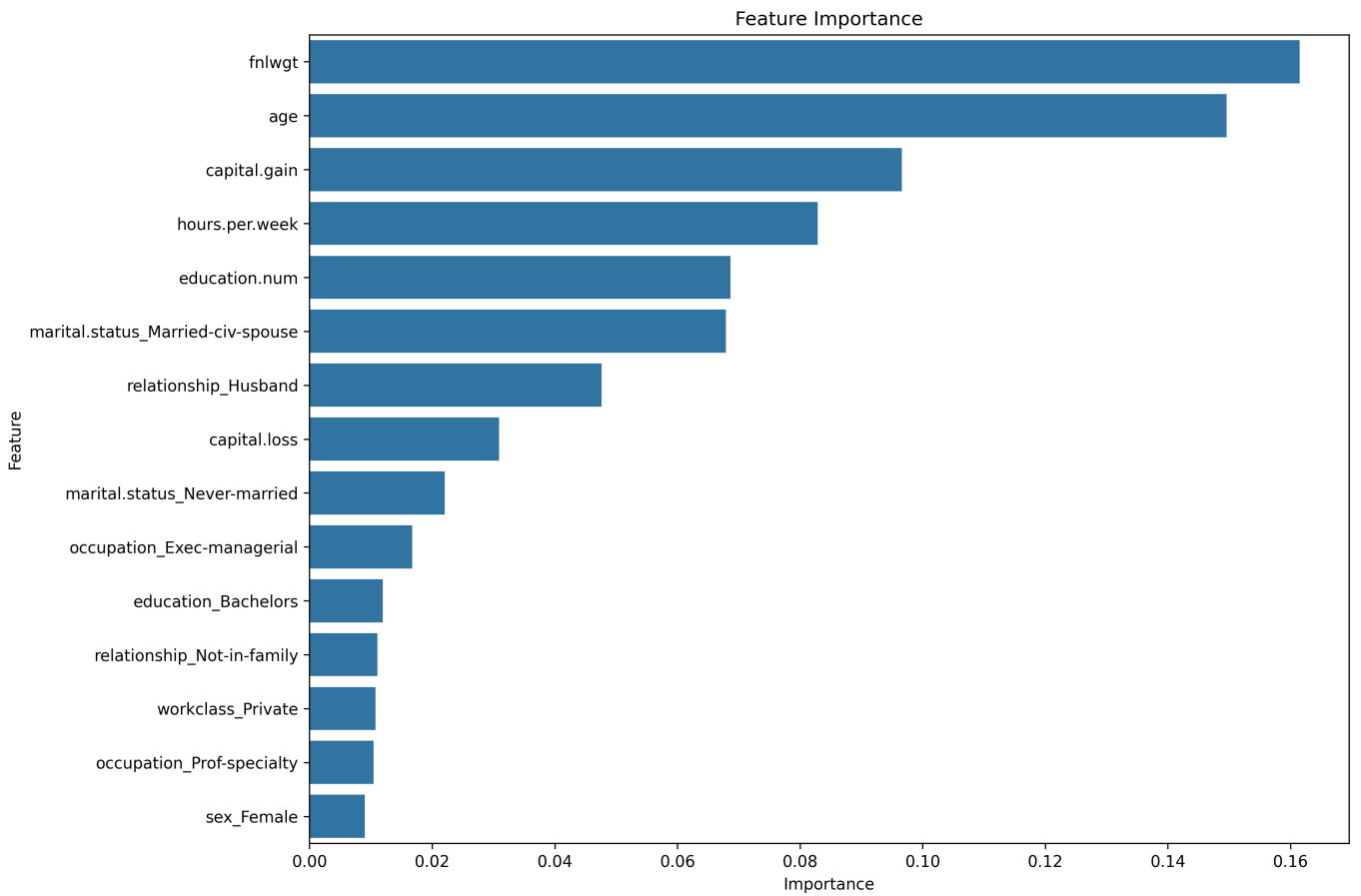


Konfusionsmatrix



Feature Importances

> *Metrik, wie viel ein Feature zur Vorhersage des Modells beiträgt*



LIME

Local Interpretable Model-agnostic Explanations

1. Grundprinzip:

- LIME erklärt einzelne Vorhersagen eines beliebigen ML-Modells
- Es arbeitet modellunabhängig (model-agnostic)
- Erzeugt lokale, interpretierbare Erklärungen für einzelne Vorhersagen

2. Funktionsweise im Detail:

Ausgangssituation:

- Ein trainiertes ML-Modell liegt vor
- Eine spezifische Vorhersage soll erklärt werden
- Das Original-Modell wird als “Black Box” behandelt

2. Funktionsweise im Detail:

Prozessschritte:

- Sampling um den Datenpunkt:
  - Erzeugt synthetische Samples in der Nachbarschaft des zu erklärenden Datenpunkts
  - Verwendet Perturbationen (kleine Änderungen) der Original-Features
- Gewichtung der Samples:
  - Näher liegende Samples erhalten höhere Gewichte

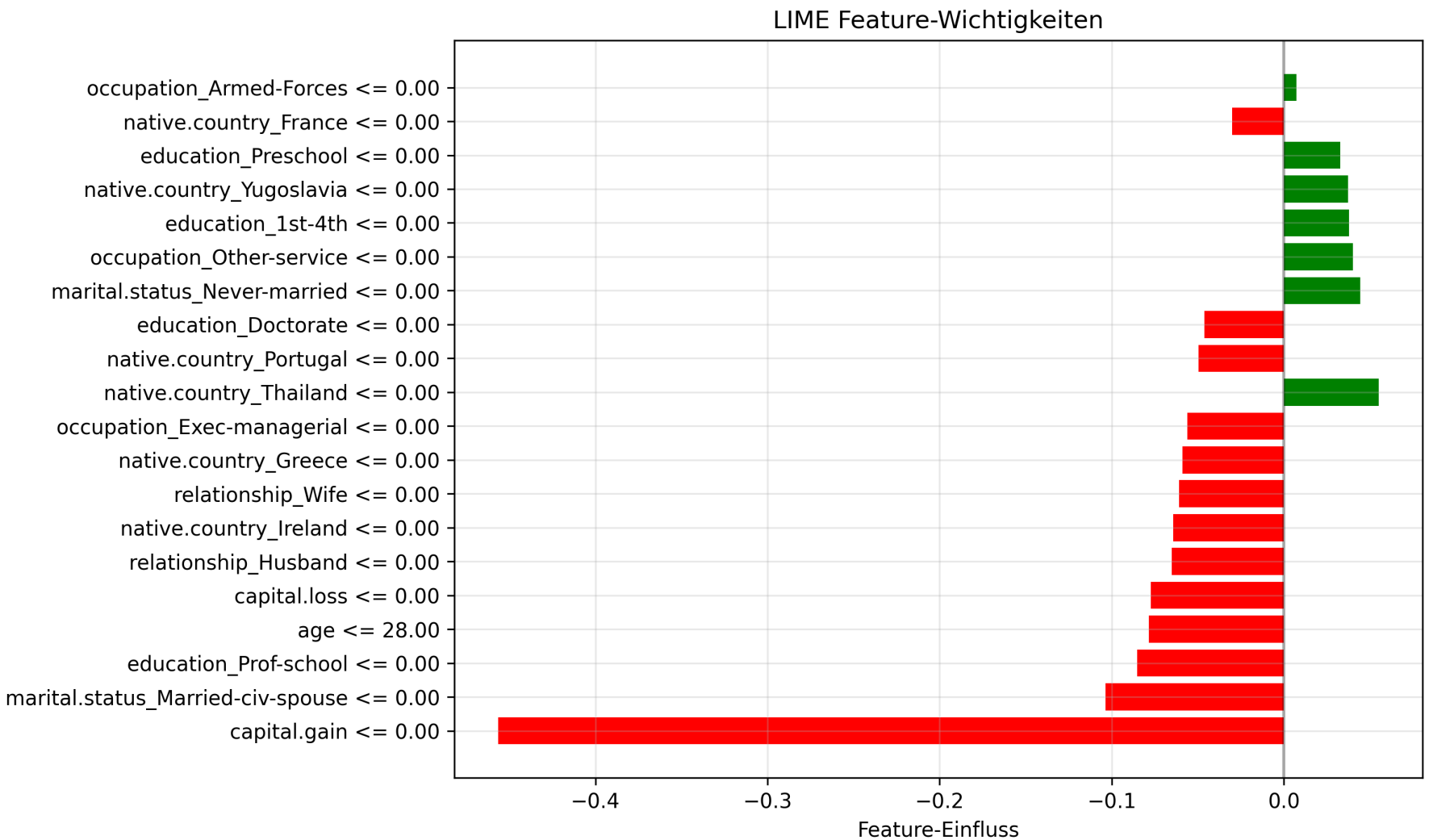
2. Funktionsweise im Detail:

Prozessschritte:

- Feature-Transformation:
  - Konvertiert die Daten in ein interpretierbares Format
  - Bei Texten z.B. Umwandlung in binäre Features (Wort vorhanden/nicht vorhanden)
- Training eines einfachen Modells:
  - typischer weise durch lineare Regression
  - Verwendet die gewichteten Samples
  - Optimiert auf lokale Genauigkeit
  - damit man ein leicht Interpretierbares Modell bekommt

2. Funktionsweise im Detail:

Extraktion der Erklärung:



3. Wichtige Eigenschaften:

- Lokalität:
  - Fokussiert sich auf lokale Umgebung der zu erklärenden Instanz
  - Erzeugt keine globalen Erklärungen für das gesamte Modell
- Interpretierbarkeit:
  - Nutzt einfache, verständliche Modelle für Erklärungen
  - Meist lineare Modelle oder Entscheidungsbäume
- Modell-Agnostik:
  - Funktioniert mit jedem ML-Modell
  - Benötigt nur Zugriff auf Vorhersagefunktion

4. Vorteile und Grenzen:

- Vorteile:
  - Flexibel einsetzbar
  - Intuitiv verständliche Erklärungen
  - Unterstützt verschiedene Datentypen
- Grenzen:
  - Nur lokale Erklärungen
  - Sampling kann rechenintensiv sein
  - LIME ist instabil
- LIME ist besonders nützlich, wenn man:
  - Einzelne Vorhersagen verstehen möchten
  - Mit komplexen Modellen arbeitet
  - Modelle debuggen oder verbessern will

LIME Anwendungsszenarien

Medizinische Diagnostik und Gesundheitswesen

- Krebsdiagnose: Erklärung, welche Merkmale in medizinischen Bildern zu einer Krebsdiagnose beitragen

LIME Anwendungsszenarien

Medizinische Diagnostik und Gesundheitswesen

- Krebsdiagnose: Erklärung, welche Merkmale in medizinischen Bildern zu einer Krebsdiagnose beitragen

Finanzwesen und Kreditvergabe

- Kreditwürdigkeitsprüfung: Transparente Begründung für Kreditablehnungen oder -genehmigungen
- Betrugserkennung: Erklärung, warum bestimmte Transaktionen als verdächtig eingestuft werden

LIME Anwendungsszenarien

Medizinische Diagnostik und Gesundheitswesen

- Krebsdiagnose: Erklärung, welche Merkmale in medizinischen Bildern zu einer Krebsdiagnose beitragen

Finanzwesen und Kreditvergabe

- Kreditwürdigkeitsprüfung: Transparente Begründung für Kreditablehnungen oder -genehmigungen
- Betrugserkennung: Erklärung, warum bestimmte Transaktionen als verdächtig eingestuft werden

Personalabteilung und Recruiting

- Bewerberselektion: Erklärung, welche Qualifikationen oder Fähigkeiten bei der Kandidatenauswahl entscheidend waren

LIME Berechnungskosten

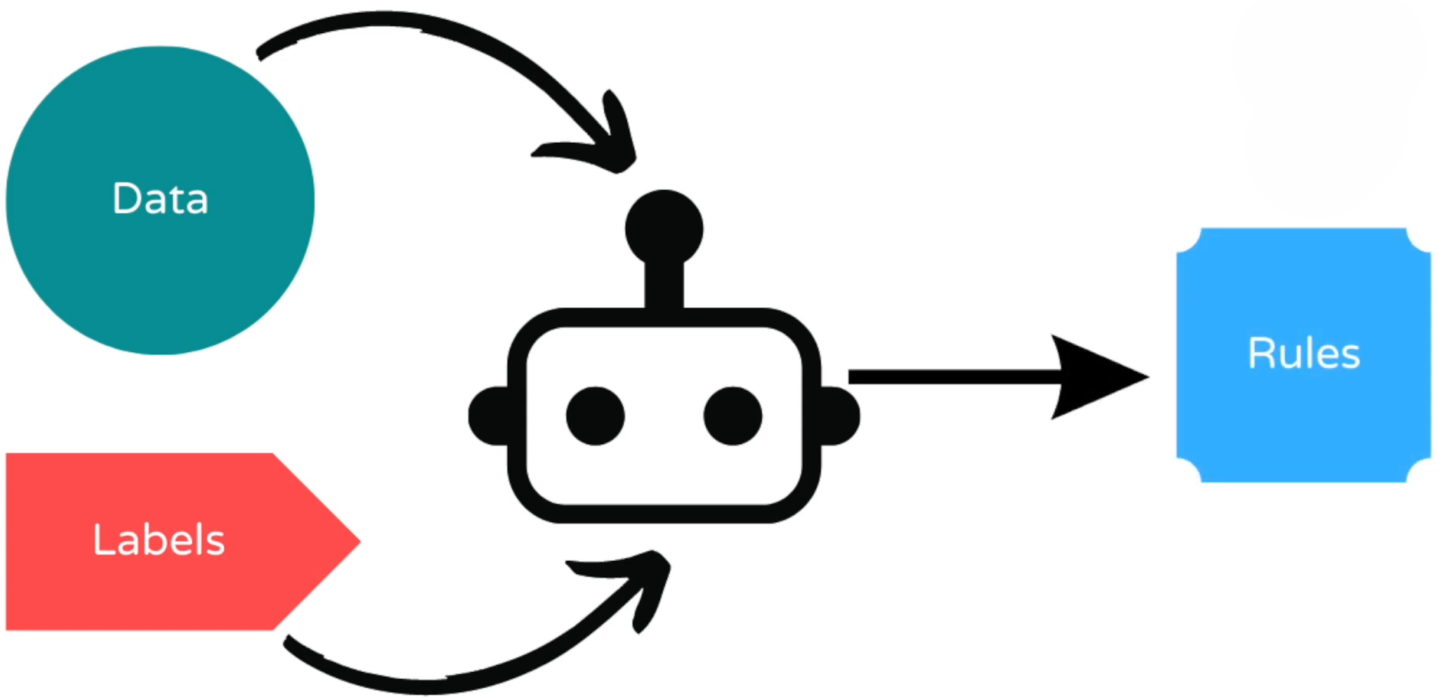
LIME skaliert:

- Linear mit der Anzahl der Samples (N)
- Quadratisch bis kubisch mit der Anzahl der Features (D)
- Linear mit der Komplexität des zu erklärenden Modells M(D)

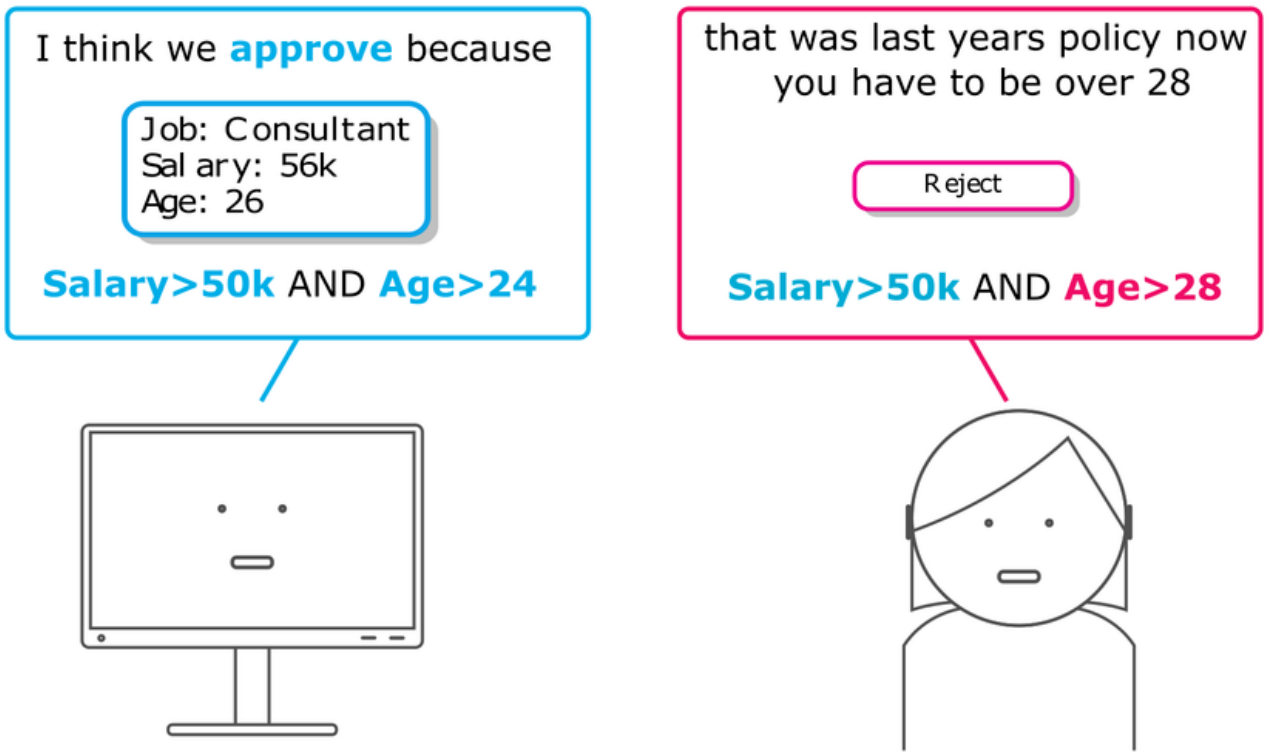
Komplexität in BigO Notation:

- **$O(N \times M(D))$**, wenn das zu erklärende Modell komplex ist
- **$O(N \times D^2)$**, wenn das lineare Modell der rechenintensivste Teil ist

Rule Based Explenation



Detaillierte Erklärung und Funktionsweise



Beispiel Regel

> WENN

> Fieber über 38°C

> UND Husten

> DANN Grippe

Beispiel Regel

> WENN
>
> capital.gain ≤ 7073.50 UND
>
> education.num ≤ 12.50 UND
>
> capital.loss > 2218.50 UND
>
> DANN Einkommen > 50K

Vorteile

- Transparenz: Menschen können Entscheidungen leicht nachvollziehen
- Überprüfbarkeit: Regeln können durch Experten überprüft und Ausgewertet werden
- Anpassbarkeit: Vordefinierte Regeln von Experten können mit maschinell erstellten Regeln ergänzt werden

Nachteile/Herausforderungen

- Skalierbarkeit: komplexen Anwendungsgebiete können eine Anzahl an Regeln schnell erreichen
- Gereralität: Regeln sind oft domänenspezifisch und benötig für unterschiedliche Kontexte entsprechend anpassungen
- Widersprüche: Regeln können zu Widersprüchen untereinar führen und müssen daher angepasst werden

RBE Anwendungsszenarien

Medizinische Diagnostik und Gesundheitswesen

- Diagnostische Systeme: Ärzte benötigen klare Regeln, um Diagnosevorschläge zu verstehen und zu validieren

RBE Anwendungsszenarien

Medizinische Diagnostik und Gesundheitswesen

- Diagnostische Systeme: Ärzte benötigen klare Regeln, um Diagnosevorschläge zu verstehen und zu validieren

Versicherungswesen

- Risikobewertung: Transparente Regeln zur Prämienberechnung
- Schadenregulierung: Nachvollziehbare Entscheidungskriterien für Schadensansprüche

RBE Anwendungsszenarien

Medizinische Diagnostik und Gesundheitswesen

- Diagnostische Systeme: Ärzte benötigen klare Regeln, um Diagnosevorschläge zu verstehen und zu validieren

Versicherungswesen

- Risikobewertung: Transparente Regeln zur Prämienberechnung
- Schadenregulierung: Nachvollziehbare Entscheidungskriterien für Schadensansprüche

Finanzwesen und Kreditvergabe

- Betrugserkennung: Einfach verständliche Regeln zur Identifikation verdächtiger Transaktionen
- Kreditentscheidungen: Banken müssen Ablehnungen oder Genehmigungen rechtlich begründen können

RBE Output

```
|--- marital.status_Married-civ-spouse <= 0.50
|   |--- capital.gain <= 7073.50
|   |   |--- education.num <= 12.50
|   |   |   |--- capital.loss <= 2218.50
|   |   |   |   |--- hours.per.week <= 40.50
|   |   |   |   |   |--- Einkommen ≤ 50K
|   |   |   |   |--- hours.per.week >  40.50
|   |   |   |   |   |--- Einkommen ≤ 50K
|   |   |   |--- capital.loss >  2218.50
|   |   |   |   |--- fnlwgt <= 125450.50
|   |   |   |   |   |--- Einkommen > 50K
|   |   |   |   |--- fnlwgt >  125450.50
|   |   |   |   |   |--- Einkommen ≤ 50K
```

RBE Berechnungskosten

RBE skaliert:

- N = Anzahl der Trainingsbeispiele
- D = Anzahl der Merkmale (Features)

Komplexität in BigO Notation: **O(N × D × log N)**

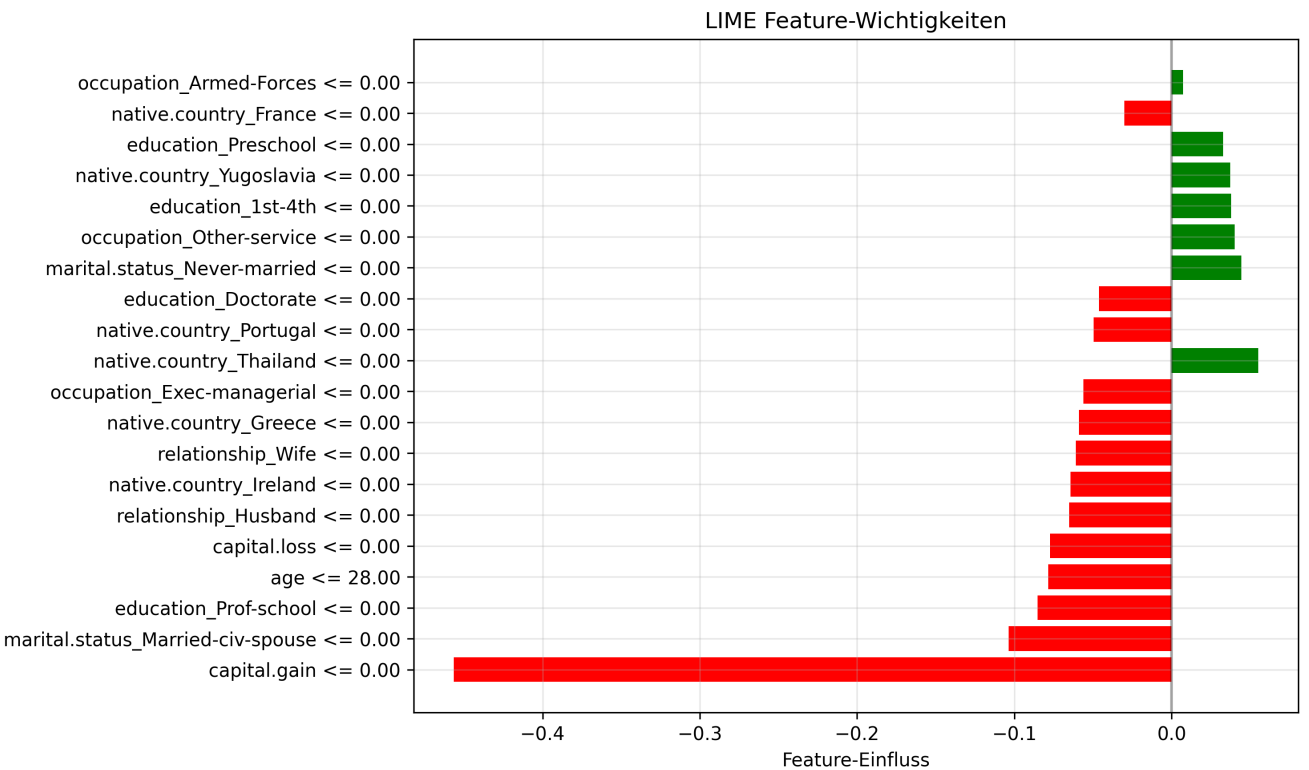
Jetzt wird es Praktisch

Vergleich der XAI-Verfahren

- Interpretierbarkeit
- Modellunabhängigkeit
- Genauigkeit / Konsistenz
- Anwendungsszenarien

LIME Interpretierbarkeit

- Leicht verständlich, welche Features relevant waren
- Auch verständlich Laien

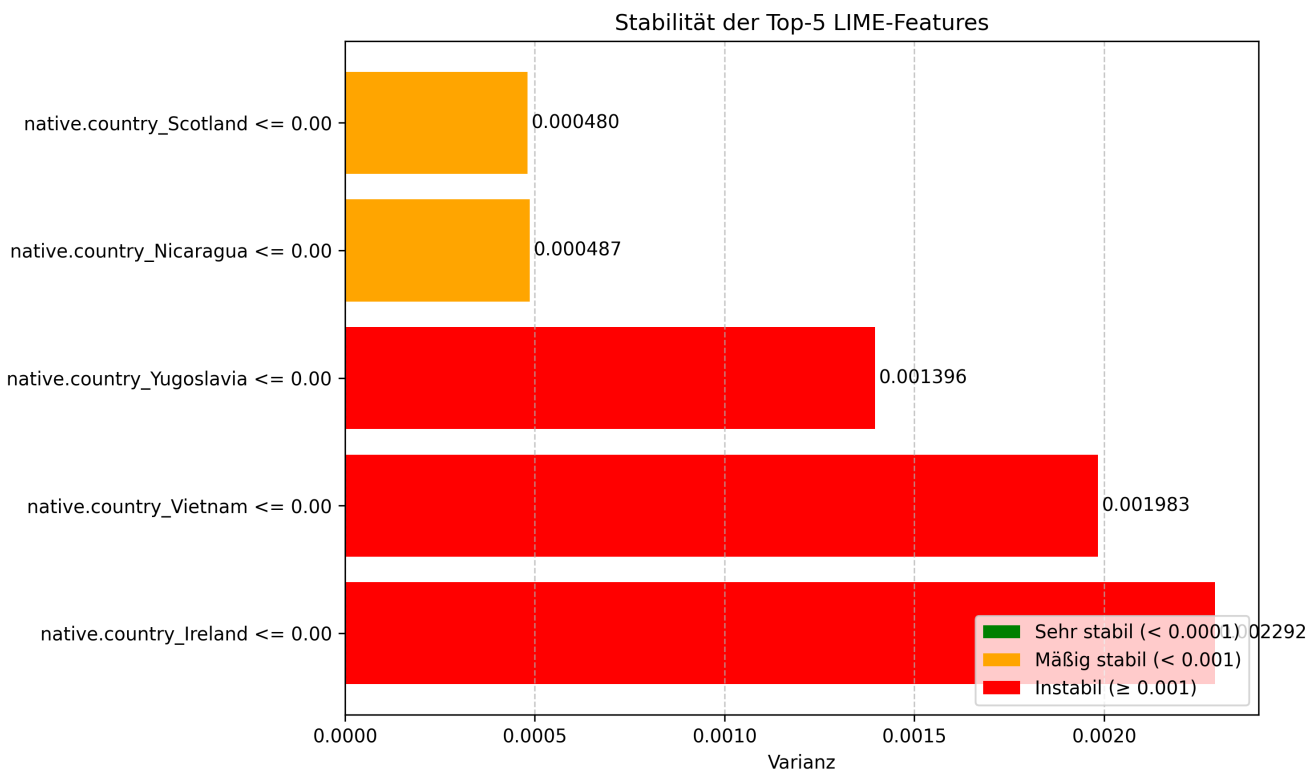


LIME Modellabhängigkeit

- Ist Model-Agnostic
- Behandelt jedes Modell als Blackbox

LIME Konsistenz

- Die Ergebnisse sind instabil
- Können aber stark variieren da es auf zufälligen Abänderungen der Daten basiert



LIME Anwendungsszenarien

- Medizinische Diagnostik und Gesundheitswesen
- Finanzwesen und Kreditvergabe
- Personalabteilung und Recruiting

RBE Interpretierbarkeit

- Regeln Leicht Nachvollziehbar: Wenn-Dann-Struktur von Regeln
- Potenziell komplex bei vielen Regeln
- Auch für Laien Verständlich

RBE Interpretierbarkeit

Beispiel einer Regel: Vorhersagt > 50K

- *WENN*
- *capital.gain ≤ 7073.50 UND*
- *education.num ≤ 12.50 UND*
- *capital.loss > 2218.50 UND*
- *DANN Einkommen > 50K*

RBE Modellabhängigkeit

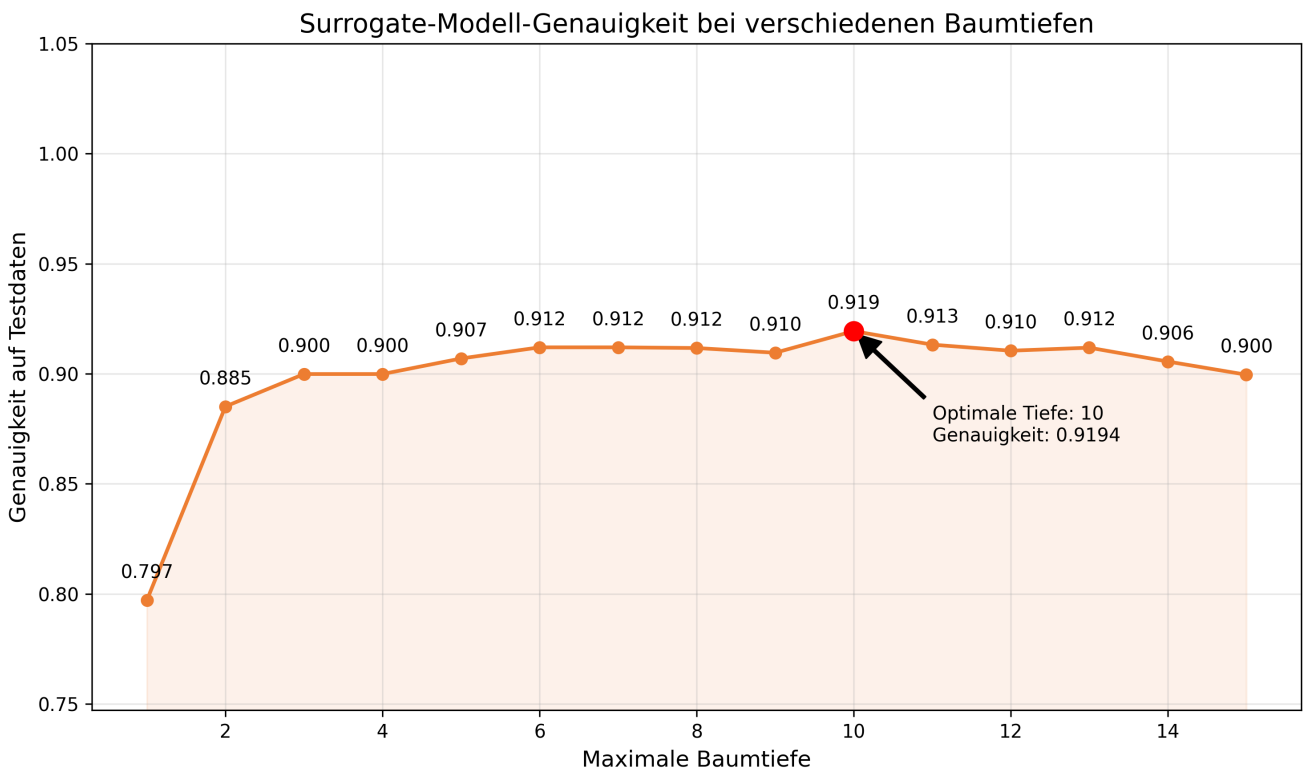
- **Inhärent regelbasierte Modelle** (modellabhängig)
- **Post-hoc regelbasierte Erklärungen** (modellunabhängig)

Wir verwenden einen: **Surrogate-Entscheidungsbaum**

- Dieser ist modellunabhängig

RBE Genauigkeit & Konsistenz

- Genauigkeit: Wie nahe an Ergebnis von RF
- Konsistenz: Schwankungen der Genauigkeit



RBE Anwendungsszenarien

- Medizinische Diagnostik und Gesundheitswesen
- Finanzwesen und Kreditvergabe
- Versicherungswesen

Vielen Dank für Eure Aufmerksamkeit!