IDA Projekt 1:Income

by Anthony Fernando

Content

- Daten Analyse
- Methoden/Workflow
  - Data Pre-processing/Vorverarbeitung
  - Feature Selection
  - Evaluation Metriken
  - Model Train/ Tune/ Validierung/Vorhersage
- Ergebnisse/Diskussion

Daten Analyse

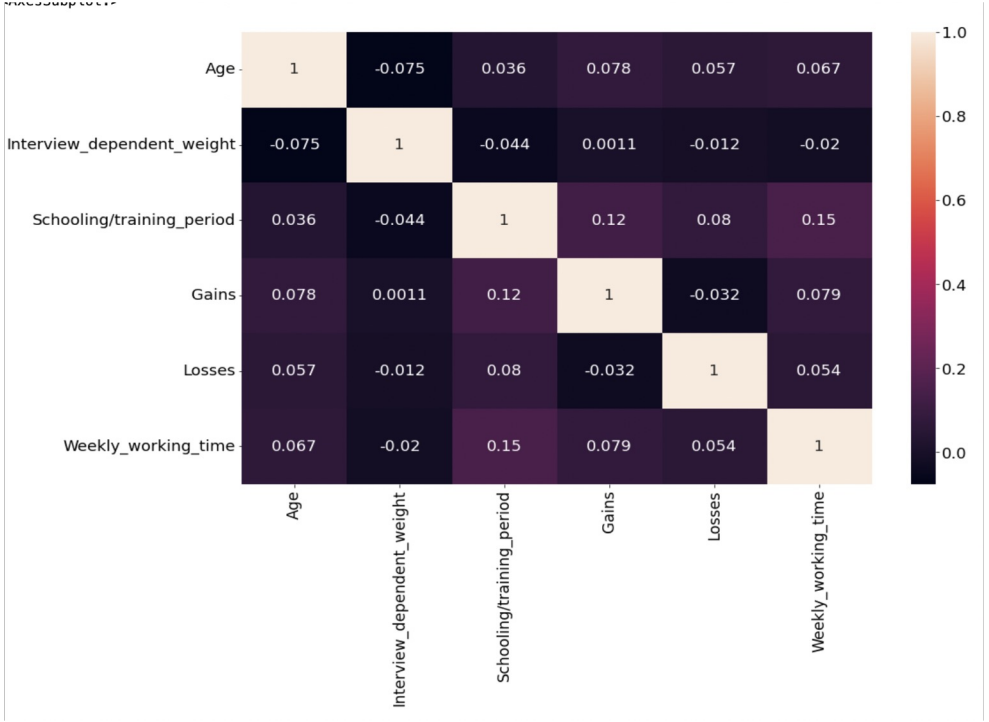
Allgemein

- 30.000 Instanzen und 15 Features
- Numerische und Kategorische Daten
- Missing values/NA:
  - Employment-type: 1677
  - Employment-area: 1682
  - Country: 539
  - Income : 25000
  - 25420 Instanzen mit mind. einen NA

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 15 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Age                         30000 non-null  int64
 1   Employment_type             28323 non-null  object
 2   Interview_dependent_weight  30000 non-null  int64
 3   Education_level             30000 non-null  object
 4   Schooling/training_period   30000 non-null  int64
 5   Marital_Status              30000 non-null  object
 6   Employment_area             28318 non-null  object
 7   Partnership                 30000 non-null  object
 8   Ethnicity                   30000 non-null  object
 9   Gender                      30000 non-null  object
 10  Gains                       30000 non-null  int64
 11  Losses                      30000 non-null  int64
 12  Weekly_working_time         30000 non-null  int64
 13  Country                     29461 non-null  object
 14  Income                      5000 non-null   object
dtypes: int64(6), object(9)
memory usage: 3.4+ MB
```

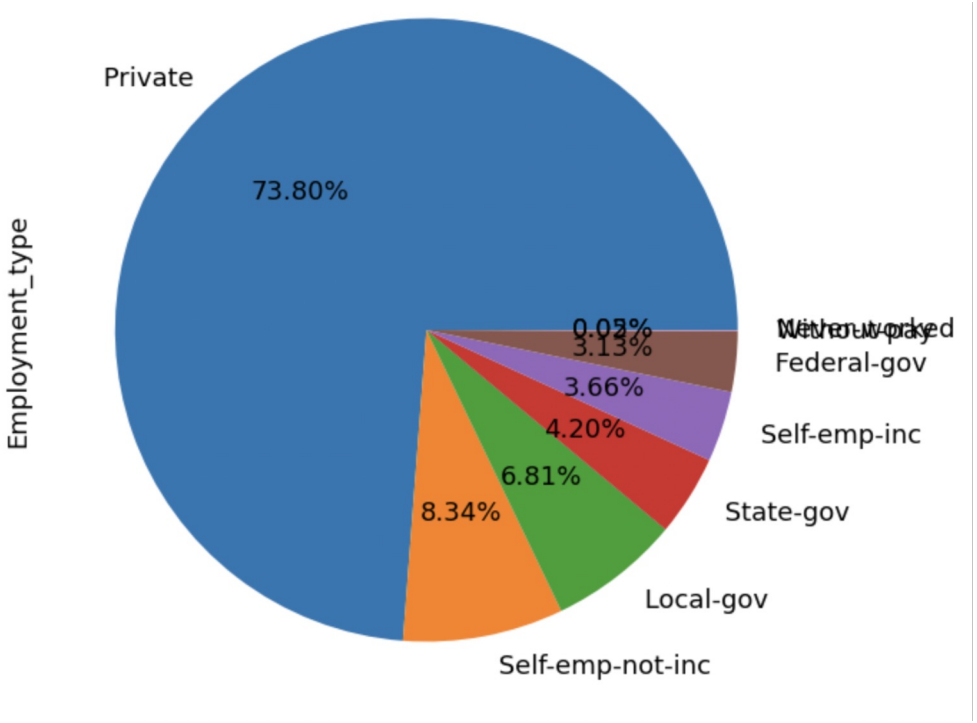
Korrelations-plot für alle Numerischen Features

- Keine große (positive) Korrelation
- Maximale Korrelation zwischen
  - Weekly working time und Schooling/Trainings (0.15)
  - Gains und Schooling/Trainings (0.12)



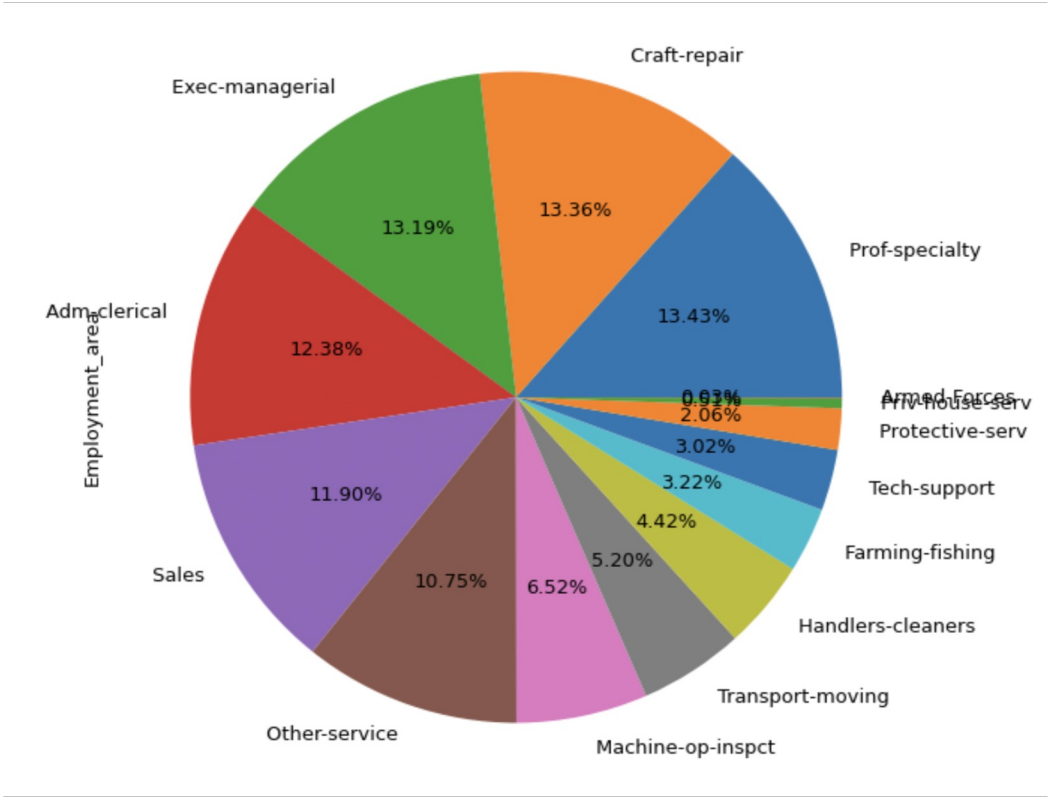
Employment Type

Private	20901
Self-emp-not-inc	2363
Local-gov	1928
State-gov	1189
Self-emp-inc	1037
Federal-gov	887
Without-pay	13
Never-worked	5



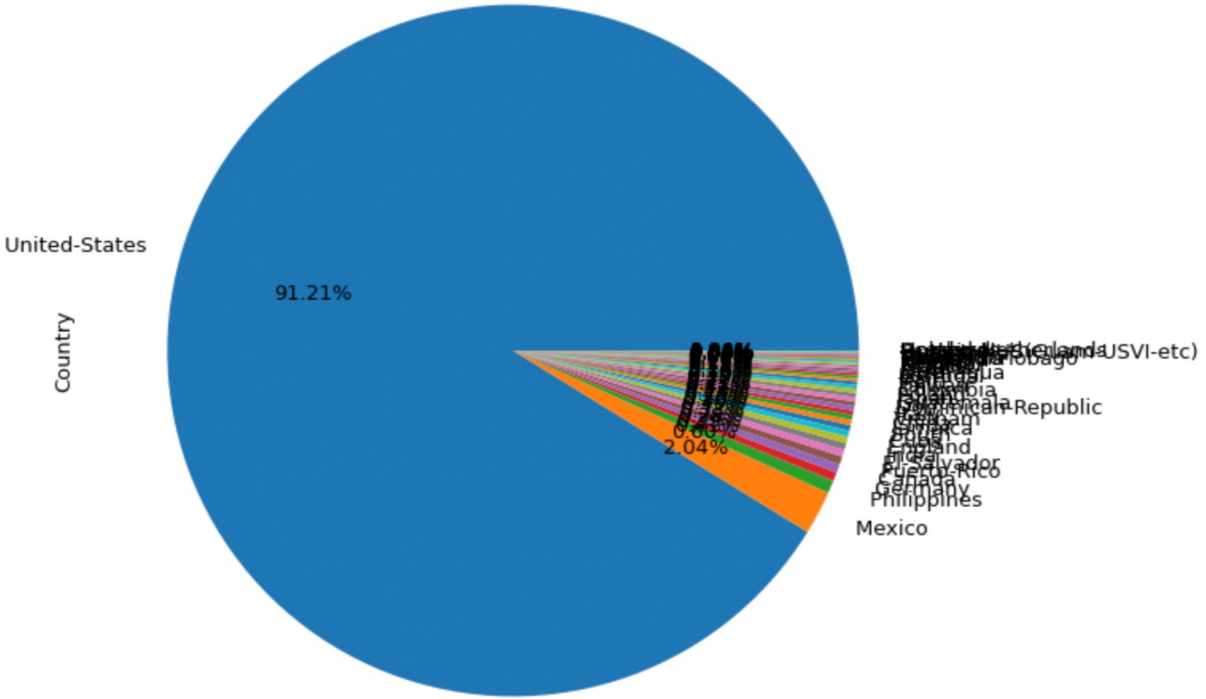
Employment Area (Beschäftigungsbereich)

Prof-specialty	3802
Craft-repair	3784
Exec-managerial	3736
Adm-clerical	3507
Sales	3370
Other-service	3044
Machine-op-inspct	1846
Transport-moving	1473
Handlers-cleaners	1252
Farming-fishing	913
Tech-support	856
Protective-serv	583
Priv-house-serv	144
Armed-Forces	8



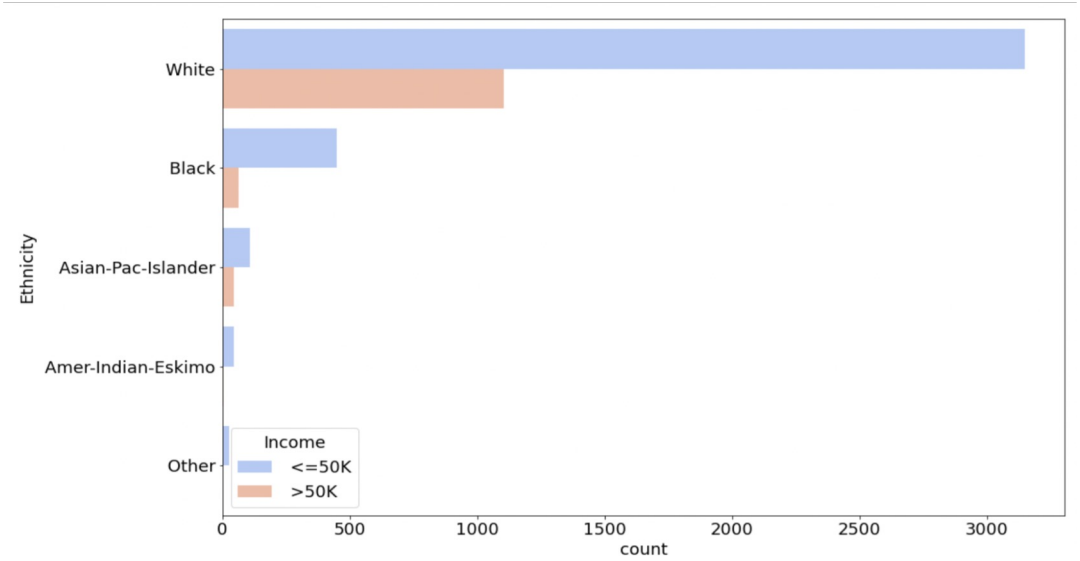
Country

- mehr als 90% stammen aus den USA
- -> der Rest ist mit 2% oder weniger vertreten
- -> Anzeichen für ein Imbalance Feature



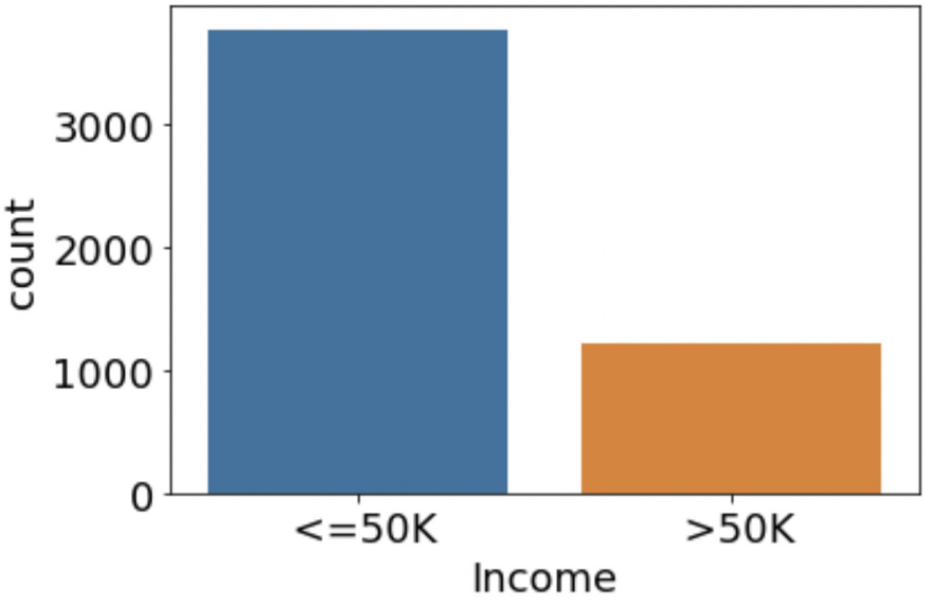
Ethnicity

- “White“ ist am häufigsten vorhanden
- 2. häufigster Wert mit nur 9 % ist „Black“
- ähnliche Verteilung wie beim Income die mehr oder weniger als 50 K verdienen

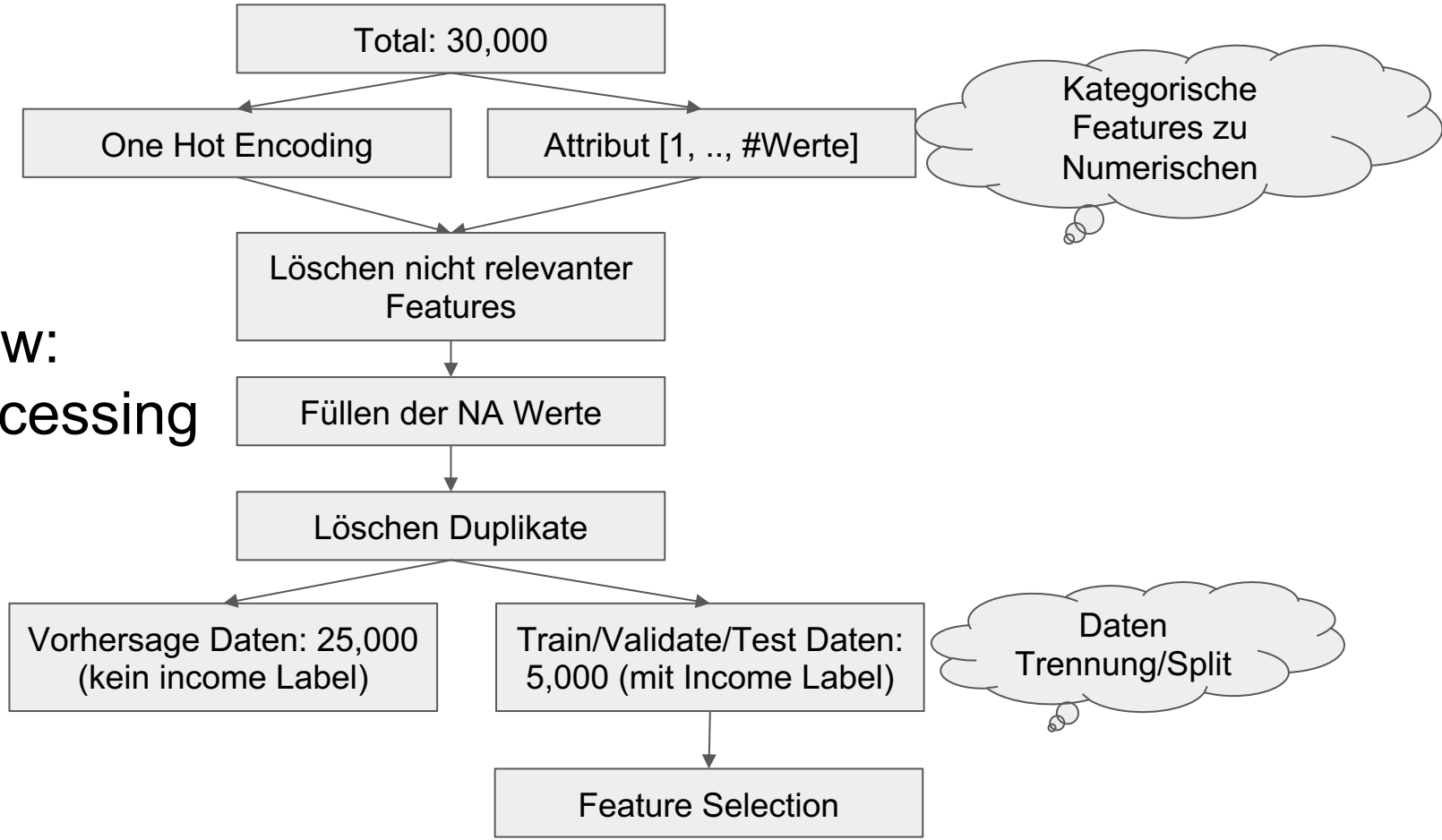


Income

- Etwas 2/3 der Samples verdienen weniger als 50 k.
- ein 1/3 davon verdienen mehr als 50 k.
- ->also moderates Class Imbalance vorhanden

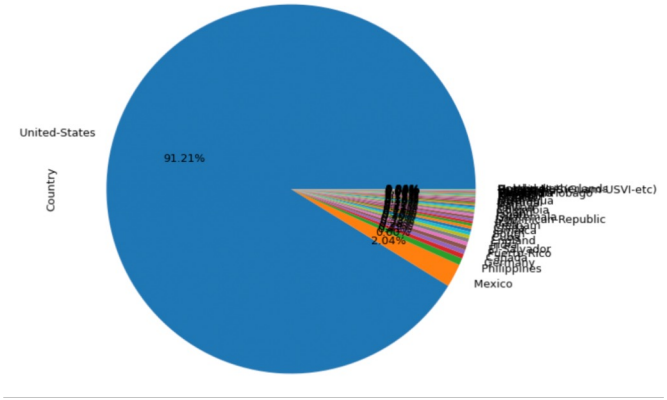


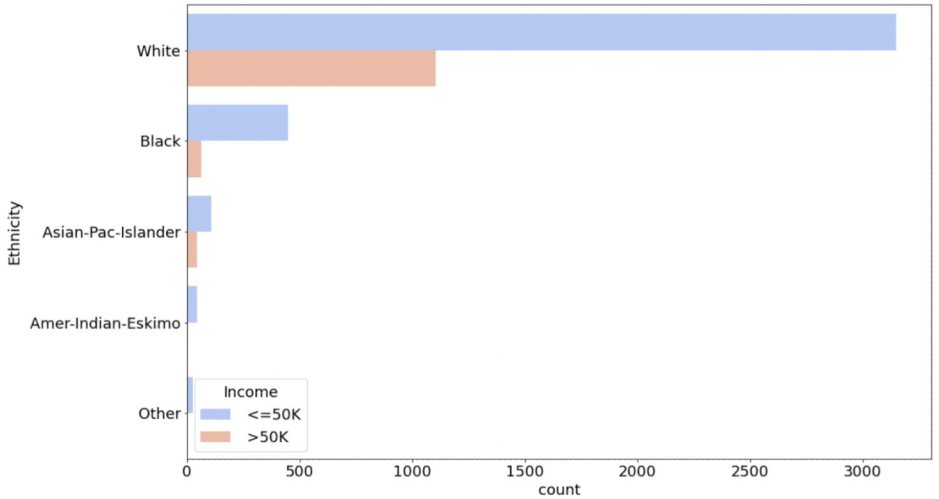
Workflow: Pre-processing



Pre-processing/Datenvorverarbeitung: Löschen nicht relevanter Features/Attribute

- Ethnicity
  - Diskriminierung
- Country
  - mehr als 90 % aus USA
  - Unausgeglichenes Feature

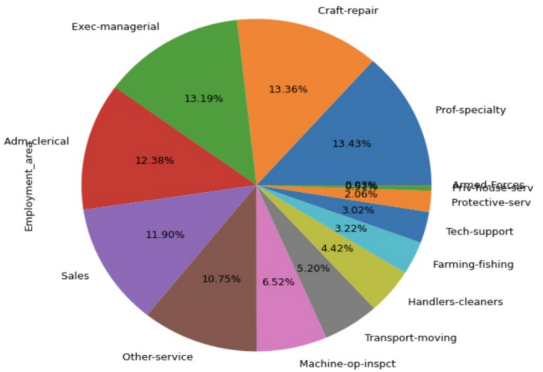




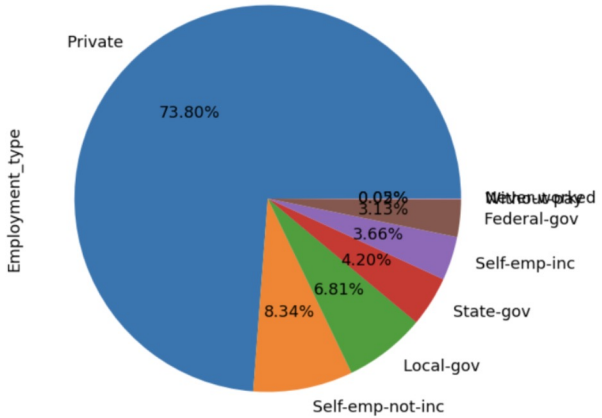
Pre-processing: Füllen NA

- Employment_type
  - Max Wert
  - ->Die meisten Instanzen haben „private“ als employment type
- Employment_area
  - Median Wert
  - Werte sind „fast“ gleichmäßig verteilt
  - Mean ungeeignet für Kategorischen werten

Prof-specialty	3802
Craft-repair	3784
Exec-managerial	3736
Adm-clerical	3507
Sales	3370
Other-service	3044
Machine-op-inspct	1846
Transport-moving	1473
Handlers-cleaners	1252
Farming-fishing	913
Tech-support	856
Protective-serv	583
Priv-house-serv	144
Armed-Forces	8



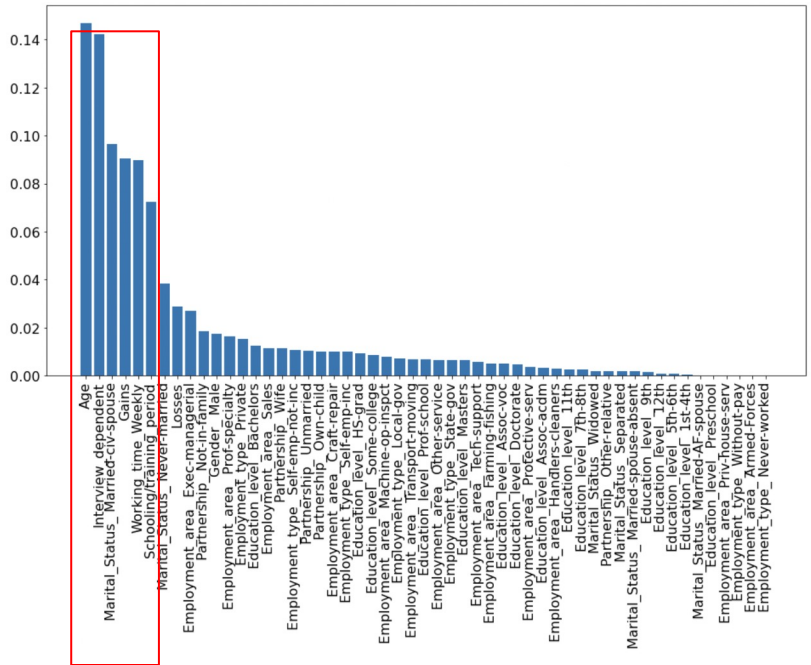
Private	20901
Self-emp-not-inc	2363
Local-gov	1928
State-gov	1189
Self-emp-inc	1037
Federal-gov	887
Without-pay	13
Never-worked	5



Feature selection : Tree.Feature Importances

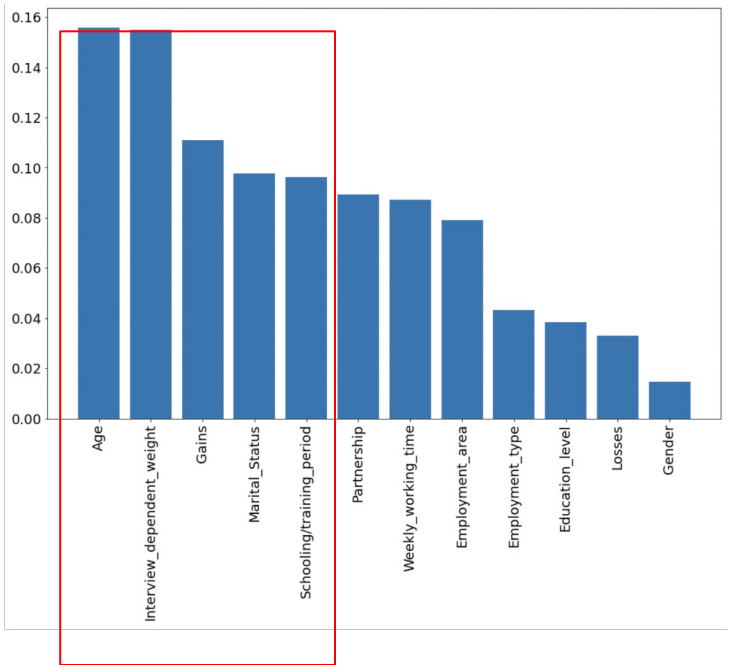
Besten 6 Features

One Hot Encoding



Besten 5 Features

Attribut [1, .., #Werte]



Ausgesuchte ML Modelle

- **Random Forest**
  - Hohe Performance beim klassifizieren und vorhersagen, selbst mit wenig Daten(unser Datensatz)
  - HP tune nicht so relevant
  - Kann mit diskreten und kontinuierlichen werten arbeiten (unser Datensatz)
  - Ersatz für Neuronale Netze
- **Decision Tree**
  - Kann mit diskreten und kontinuierlichen werten arbeiten (unser Datensatz)
  - Einfach zu interpretieren
  - Vergleich zwischen 1 Baum vs. Mehrer Bäume
- **SVM Classifier**
  - Vermutung das“ Daten linear separiert“ sind nicht mehr notwendig
  - Findet beste lineare hypertrennebene
  - Nutz Kerneltrick
  - Geringe Chance für Overfitting
- **Logistische Regression**
  - Keine Annahme über die Verteilung der Klassen (Class Imbalance egal)
  - Bases Model
  - Bei niedrig Dimensionalen Datensatz und genug Training-instanzen: Weniger anfällig für Overfitting

Hyperparameter für die jeweiligen Modelle

- Random Forest : # d. Bäume [1-100]
    - Tiefe: 4

- Decision Tree : Tiefe des Baumes[1-54]
    - Gewählt anhand der Feature Zahl
    - Criterion: Entropy

- SVM Classifier: C(Regularization parameter)[100, 10, 1.0, 0.1, 0.001]
    - Gamma: Auto

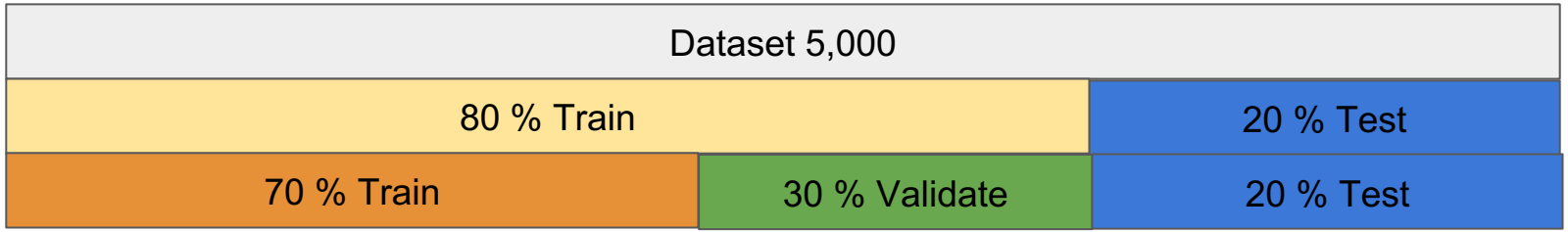
- Logistische Regression : C(Regularization parameter)[100, 10, 1.0, 0.1, 0.001]
    - L2 Regularisierer

-> es wurde nur ein Parameter getuned um fairen vergleich zu machen
-> nicht bei allen Modellen gab es immer 2 eindeutige Parameter
-> Parameter für Random Forest, SVM und Logistische Regression basieren auf Literatur suche

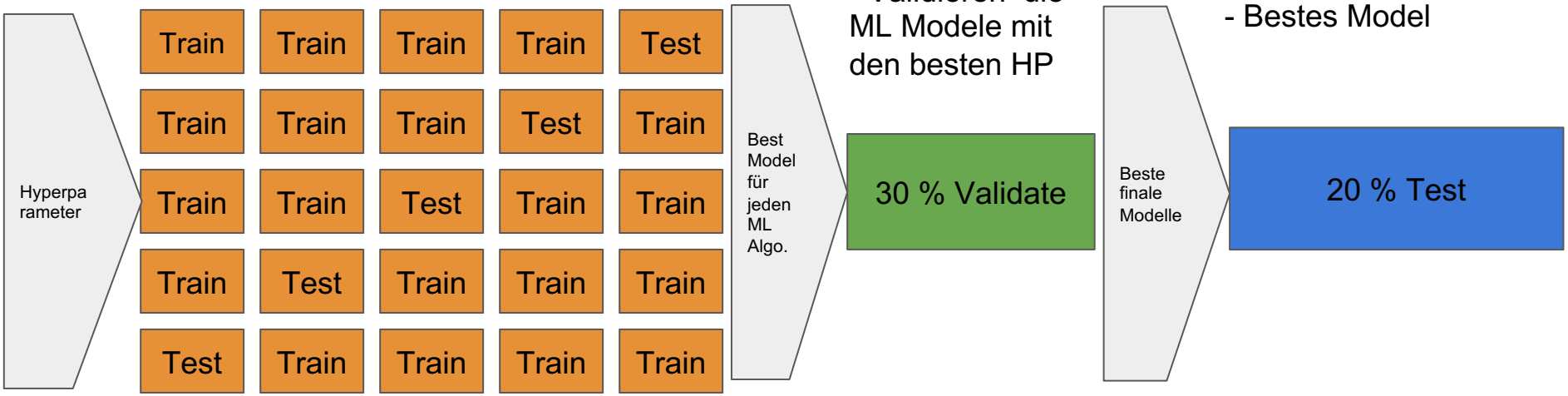
Evalutions Metriken

- Accuracy
  - Mittelwert und Standardabweichung
  - -> SD um zu sehen wie stark sie vom Mittelwert streuen
- F1 Score
  - Gute für unausgeglichen daten
- Konfusion Matrix
  - Um zu sehen wie viele Klassen falsch klassifiziert wurden
- k-fold cross-validation
  - k=5
  - Um Overfitting zu vermeiden
- ROC /AUC

Model Training/Tuning/Testing



- 5- fold cross validation zum tunen und trainieren der HP



- Validieren die ML Modele mit den besten HP

- Bestes Model

Result: Attribut[1, .., #Werte] mit Feature Selection

ML Model	Best Parameter	Accuracy(Mittel wert)	Accuracy(SD)
Random Forest	# d. Bäume : 50	0.830	0.005
Decision Tree	Tiefe des Baumes: 5	0.829	0.007
SVM	C: 10	0.827	0.006
Logistische Regression	C: 0.001	0.788	0.010

Bestes

Schlechtestes

Result: Attribut [1, .., #Werte] mit Feature Selection

ML Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.841	0.733	0.465	0.569
Decision Tree	0.846	0.729	0.506	0.597
SVM	0.847	0.746	0.487	0.589
Logistische Regression	0.801	0.742	0.181	0.291

30 % Validate

Attribut [1, .., #Werte]

Bestes

Schlechtestes

Result: Attribut [1, .., #Werte] ohne Feature Selection

ML Model	Best Parameter	Accuracy(Mittelwert)	Accuracy(SD)
Random Forest	# d. Bäume: 40	0.832	0.007
Decision Tree	Tiefe des Baumes: 4	0.830	0.008
SVM	C:10	0.821	0.012
Logistische Regression	C: 0.001	0.786	0.013

Bestes

Schlechtestes

Result: Attribut [1, .., #Werte] ohne Feature Selection

ML Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.845	0.764	0.454	0.569
Decision Tree	0.845	0.716	0.520	0.603
SVM	0.839	0.681	0.542	0.604
Logistische Regression	0.806	0.716	0.232	0.351

30 % Validate

Attribut [1, .., #Werte]

Bestes

Schlechtestes

Result: One-Hot-Encoding mit Feature Selection

One Hot Encoding	70 % Train

ML Model	Best Parameter	Accuracy(Mean)	Accuracy(SD)
Random Forest	# d. Bäume :98	0.834	0.007
Decision Tree	Tiefe des Baumes: 4	0.834	0.009
SVM	C: 1.0	0.829	0.008
Logistische Regression	C: 0.001	0.789	0.010

Bestes

Schlechtestes

Result One-Hot-Encoding mit Feature Selection

ML Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.845	0.732	0.494	0.590
Decision Tree	0.848	0.735	0.513	0.604
SVM	0.844	0.723	0.502	0.593
Logistische Regression	0.797	0.700	0.181	0.287

30 % Validate

One Hot Encoding

Bestes

Schlechtestes

Result: One-Hot-Encoding ohne Feature Selection

One Hot Encoding	70 % Train

ML Model	Best Parameter	Accuracy(Mean)	Accuracy(SD)
Random Forest	# d. Bäume:35	0.829	0.009
Decision Tree	Tiefe des Baumes: 4	0.832	0.009
SVM	C: 1.0	0.828	0.011
Logistische Regression	C: 0.001	0.791	0.011

Bestes

Schlechtestes

Result: One-Hot-Encoding ohne Feature Selection

ML Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.836	0.819	0.351	0.491
Decision Tree	0.847	0.721	0.524	0.607
SVM	0.858	0.734	0.579	0.647
Logistische Regression	0.803	0.684	0.240	0.355

30 % Validate

One Hot Encoding

Bestes

Schlechtestes

Result: Finales Model Test vergleich

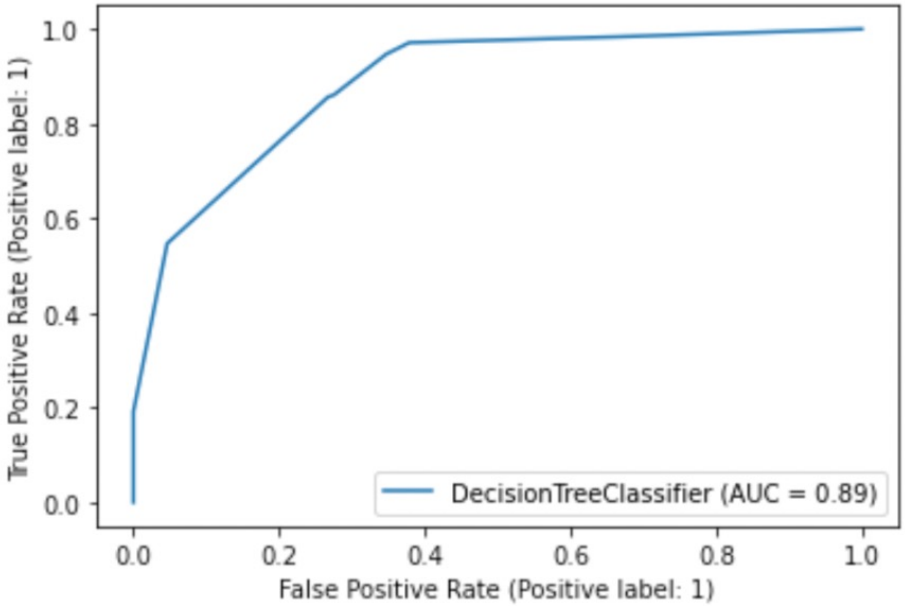
20 % Test

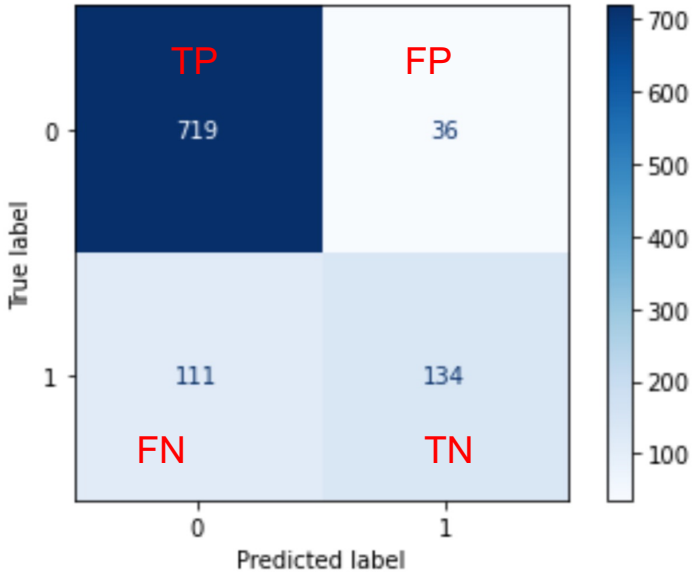
ML Model	Numeric /One hot	Feature Selection	Accuracy	Precision	Recall	F1-Score
Decision Tree	Numeric	Yes	0.853	0.788	0.547	0.646
Decision Tree	Numeric	No	0.850	0.771	0.551	0.643
Decision Tree	One Hot	Yes	0.854	0.793	0.547	0.647
SVM	One Hot	No	0.848	0.702	0.565	0.626

Bestes

Schlechtestes

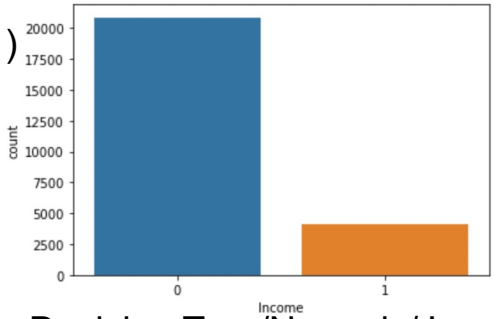
ROC Kurve für das beste Model(Model 3)





Vorhersage Ergebnisse für alle 4 Modelle

1)

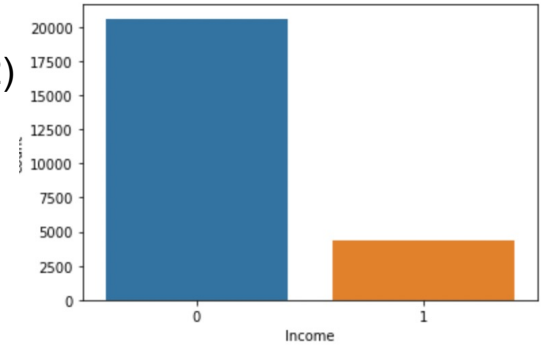


Decision Tree/Numeric/Ja

0 0.835361

1 0.164639

2)

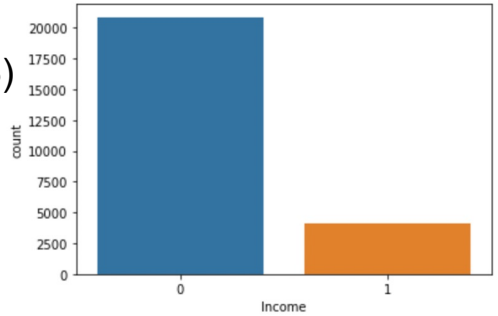


Decision Tree/Numeric/Nein

0 0.824754

1 0.175246

3)

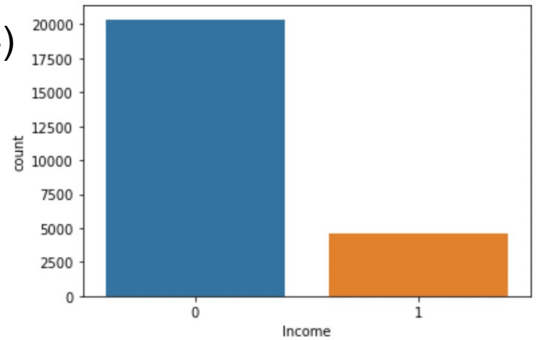


Decision Tree/One Hot/Ja

0 0.835041

1 0.164959

4)

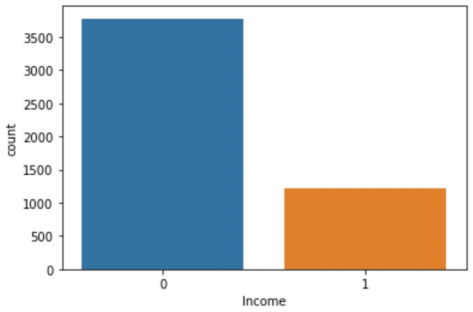


SVM /One hot/Nein

0 0.815107

1 0.184893

Original)



0 0.755751

1 0.244249