ETH zürich

Social Mobility and Education

Date Science in Techno-Socio-Economic Systems

19. May 2025

EDUCATION

- **Years** of studies
- Household **assests** available

FAMILY

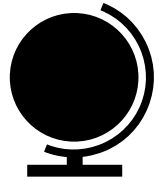
- **Highest** education
- Education **Father**
- Education **Mother**
- Number of siblings

RELIGION

- Christian, Muslim, Hindu, etc.
- Different **Caste**

INFLUENCE

- **Education** compared to other Attributes
- **Relevance**



GLOBAL

- **Intercontinental**
- Cross-country **differences**

Motivation

Open Questions

EDUCATION

Does education tell us anything about current / future income?

FAMILY

Does the environment in which a person is growing up tell us something about current / future income?

RELIGION

Is the prediction biased on the religion or culture one has.

INFLUENCE

If measurable, how much influence do different kinds of attributes (features) have on the predictions?

GLOBAL

Are there cross-country differences in mobility patterns?

GENERAL

How to find these correlations?

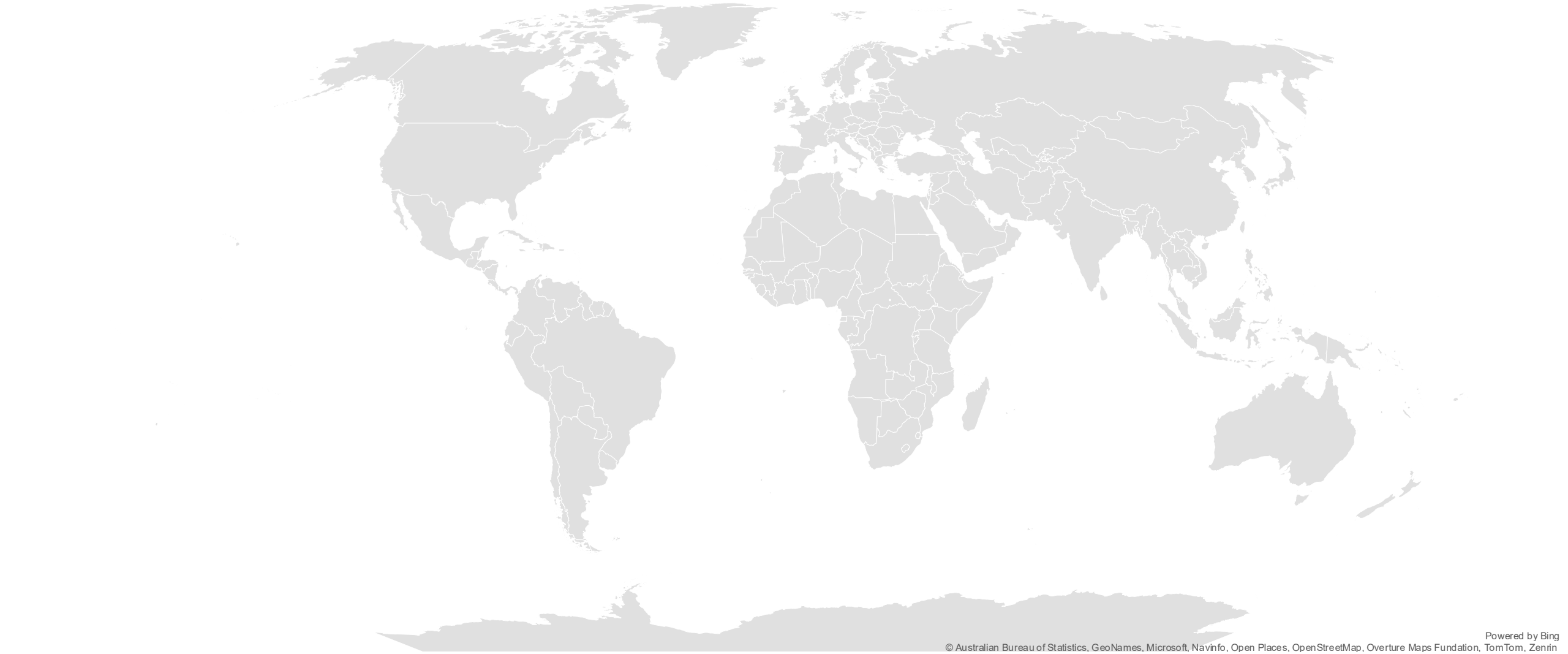
How to find these correlations?

NEURAL NETWORK
For learning highly complex, non-linear patterns

RANDOM FOREST
Captures complex, non-linear relationships

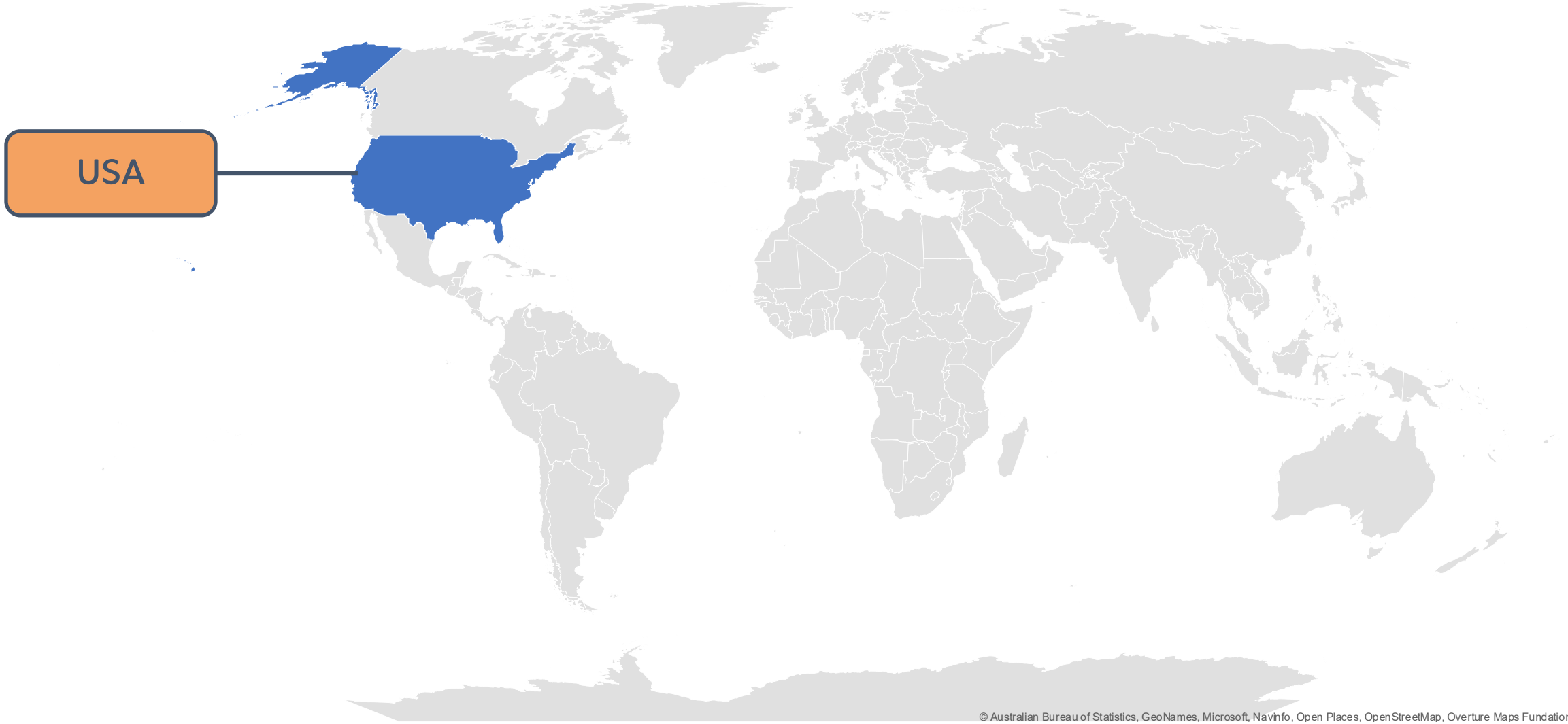
MULTIPLE LINEAR REGRESSION
Quickly interprets linear relationships

Country Analysis

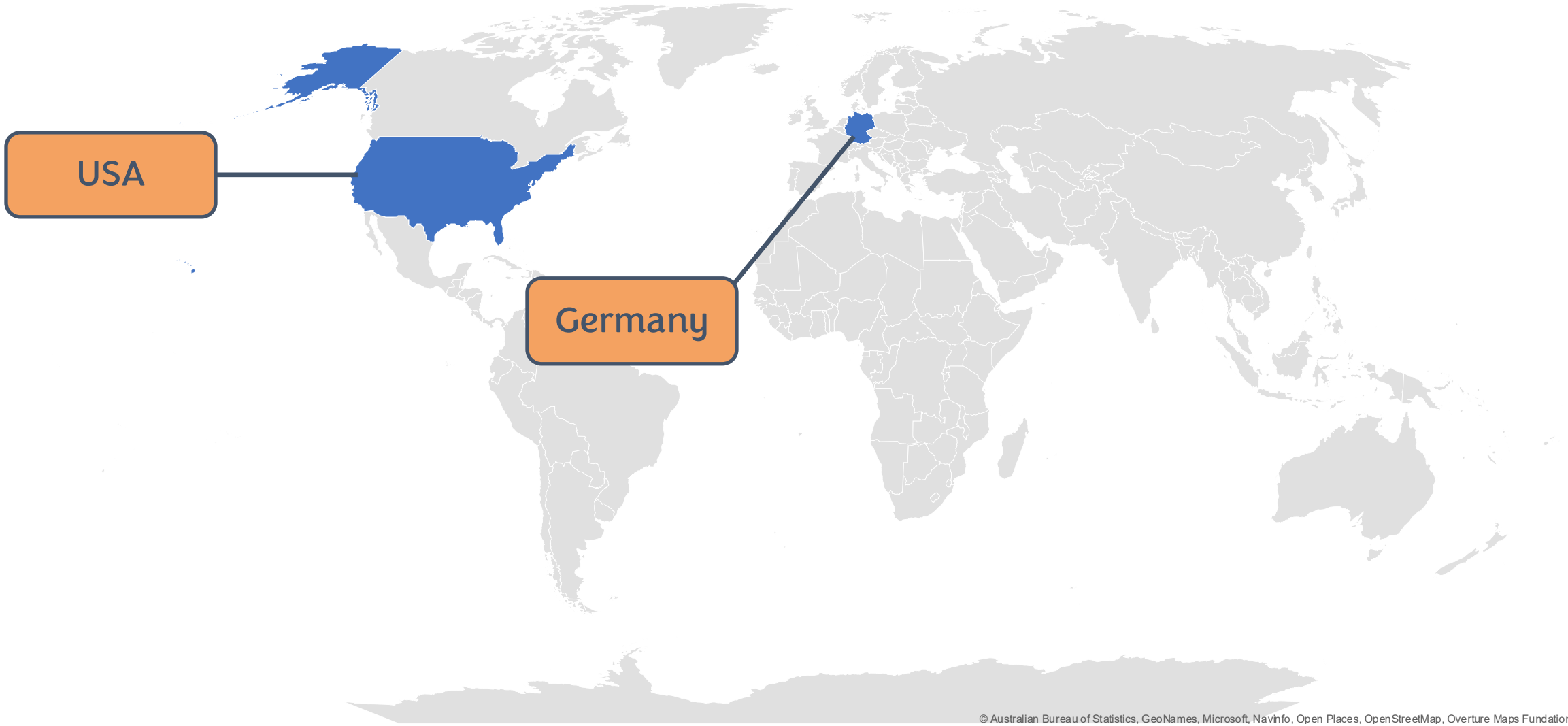


Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, Open Places, OpenStreetMap, Overture Maps Fundation, TomTom, Zenrin

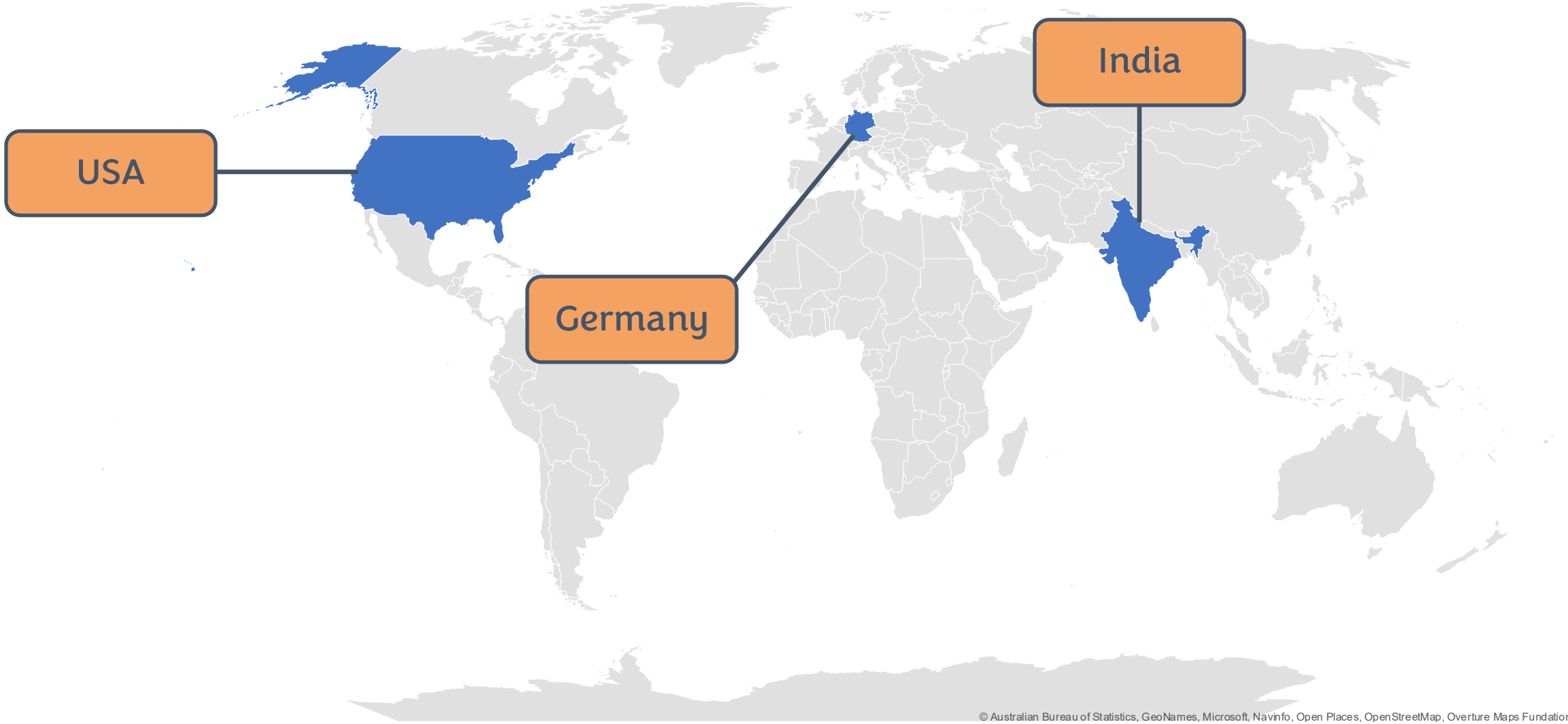
Country Analysis



Country Analysis



Country Analysis



Hypothesis

1. The combination of education level and environmental quality metrics predicts income mobility better than either factor alone.

2. Higher levels of individual education are associated with increased upward social mobility, as measured by a weaker correlation between other features such as parental income, parental education, …

USA (aggregate data)

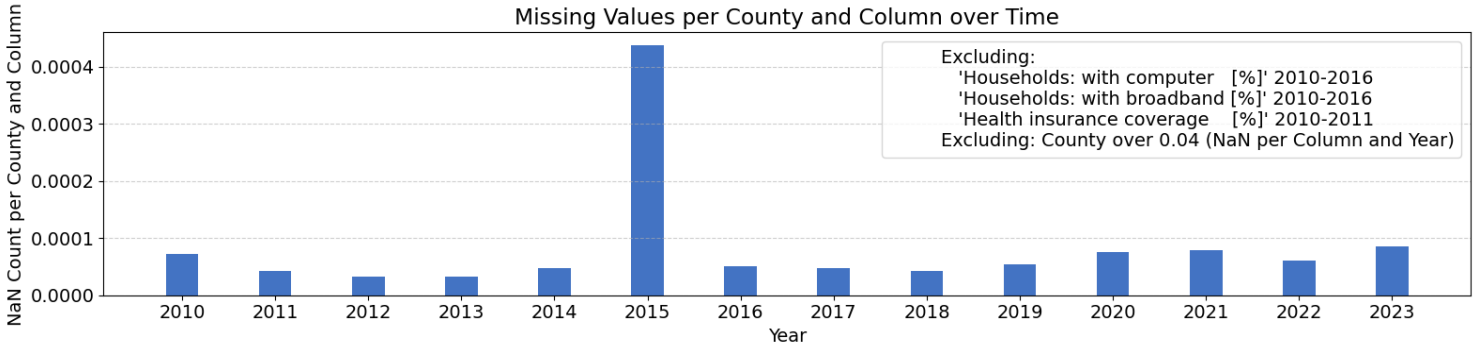
Data & Model

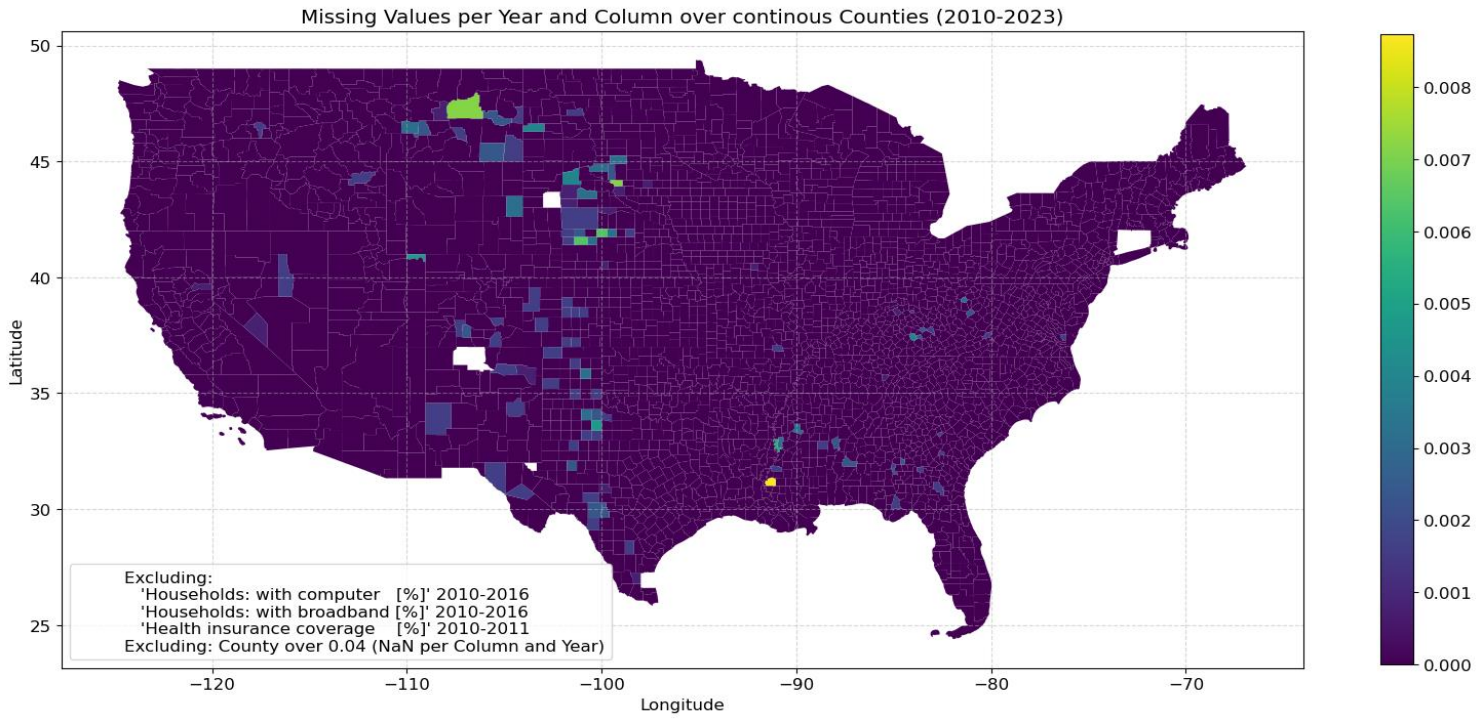
DATA

- U.S. Census Bureau:
  America Community Survey:
  5-Year Estimates Data Profiles
- U.S. Census Bureau: Tiger Shapefile
- Tables: 14 * 5 = 70
- Timeframe: 2010-2023
- Geographic Scope:
  Continious Mainland U.S. Counties

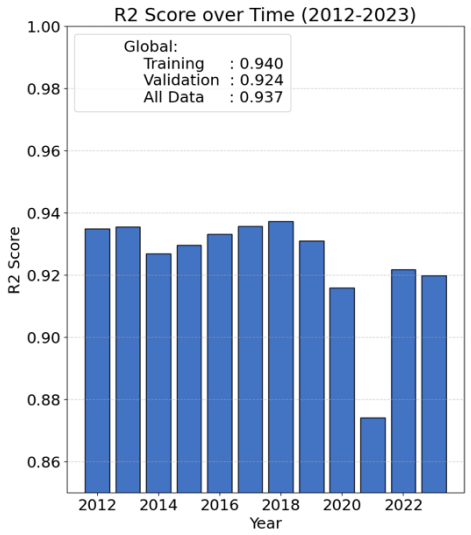
MODEL

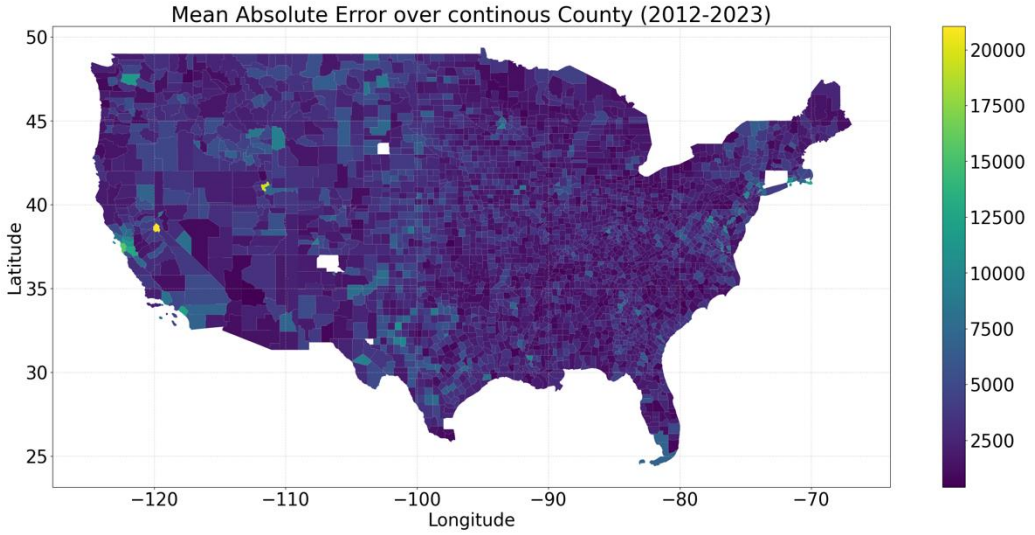
- LSTM-based Model
- High Total Feature Dropout
- 73 Features
- 3 Year Context Window

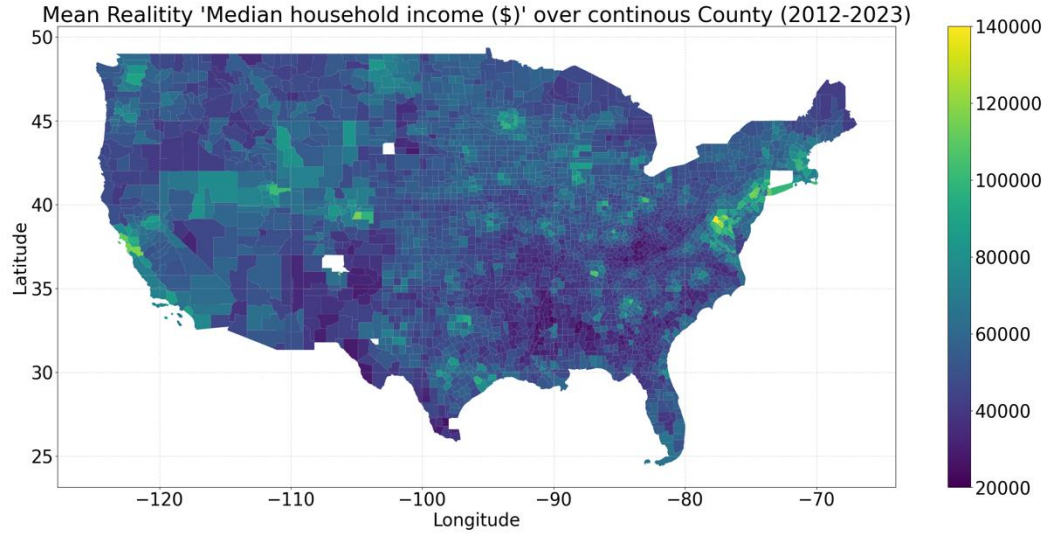


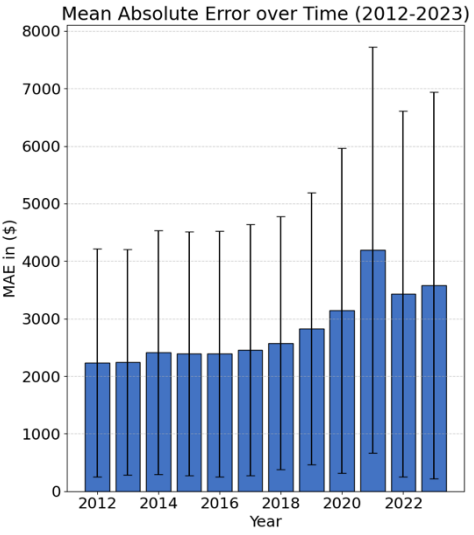


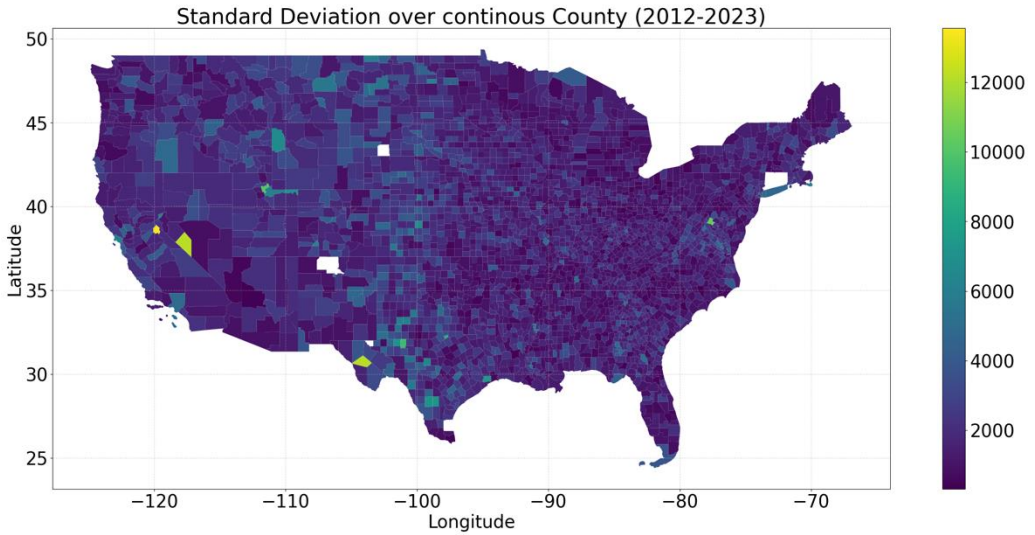
Model Quality

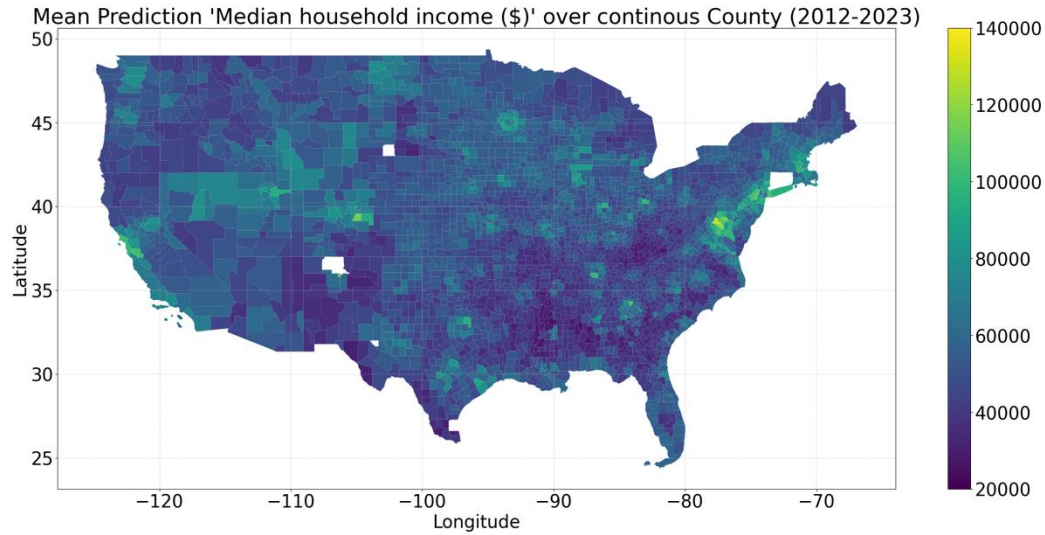




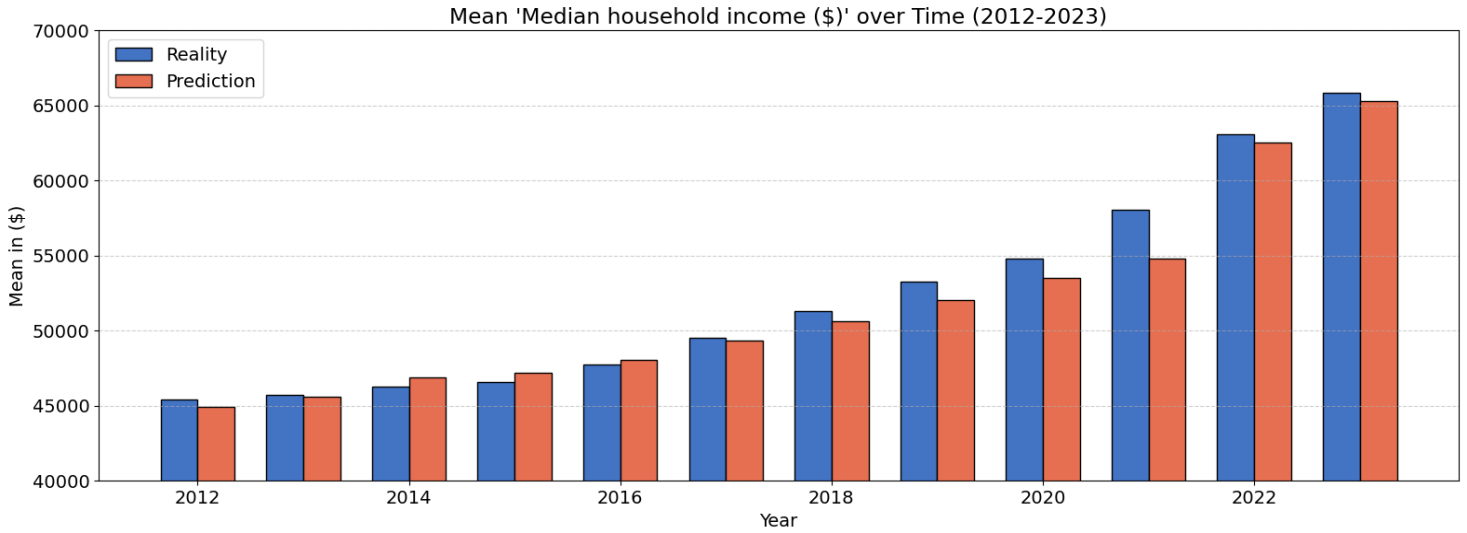


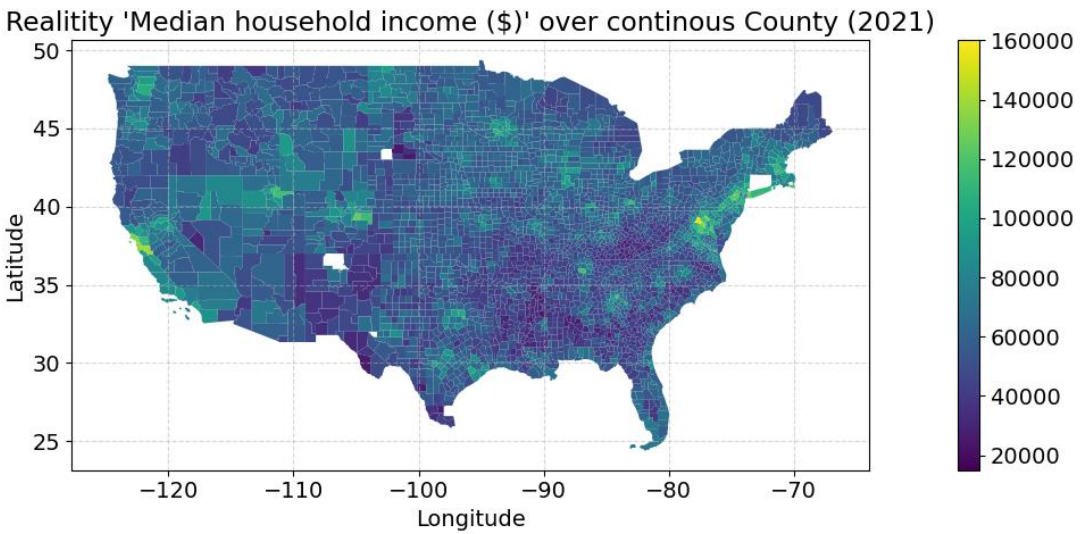


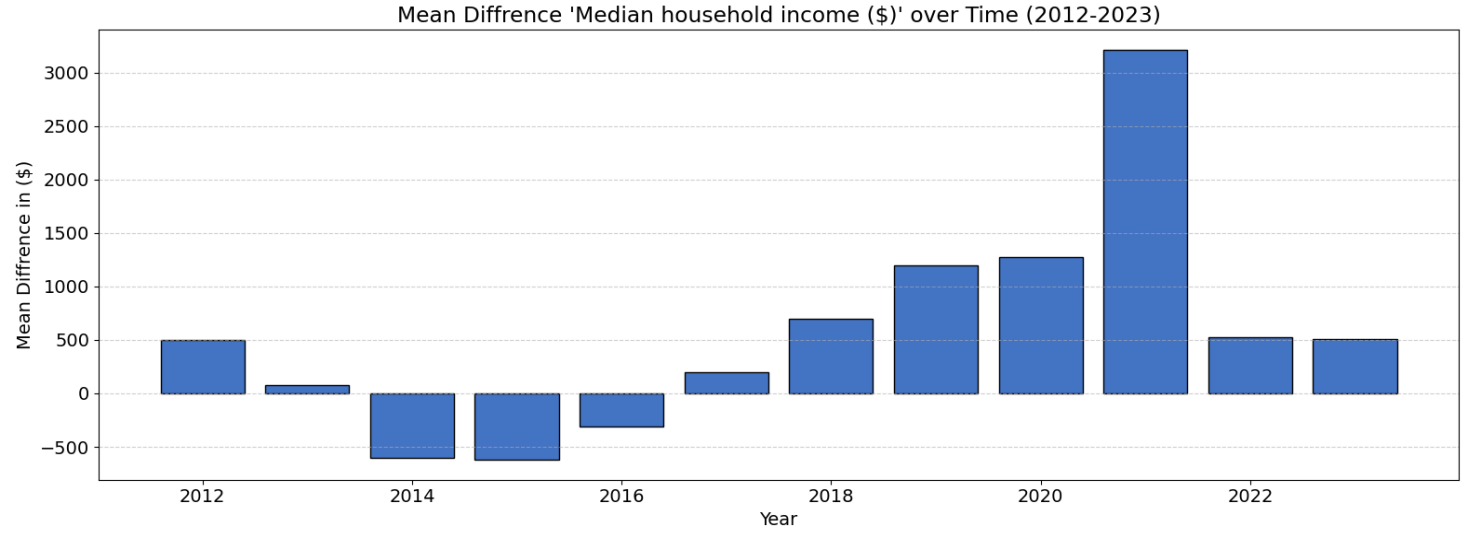


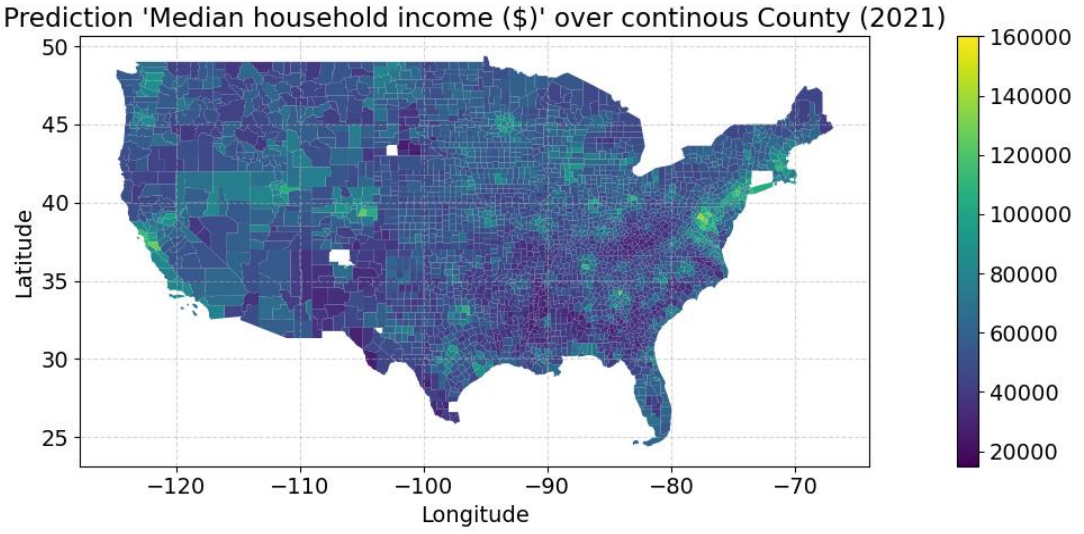


Prediction

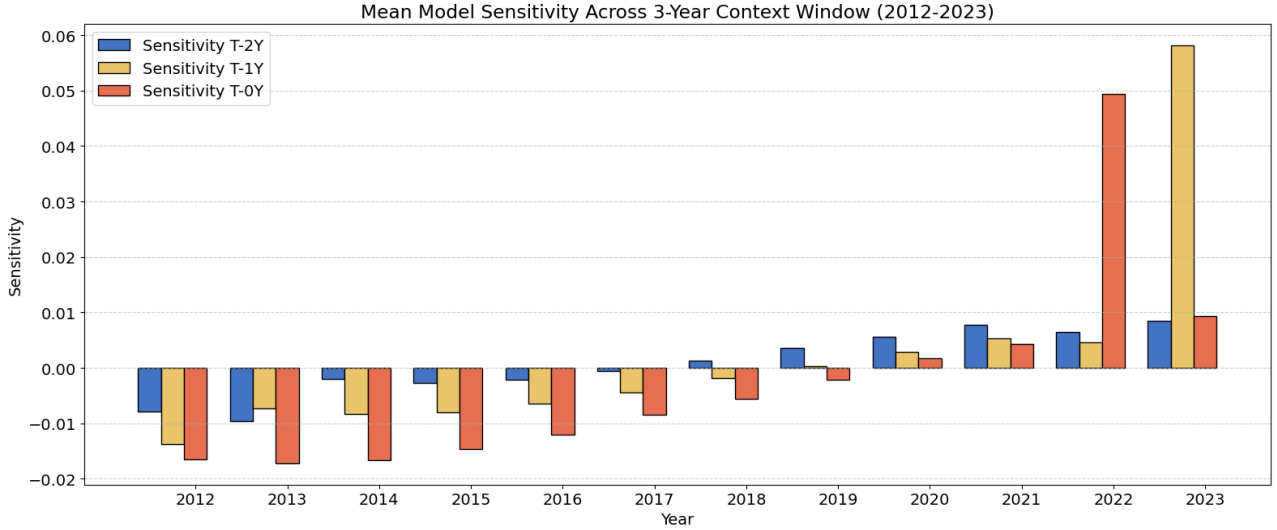

Mean 'Median household income ($)' over Time (2012-2023)
Reality
Prediction
Mean in ($)
Year
70000
65000
60000
55000
50000
45000
40000
2012
2014
2016
2018
2020
2022

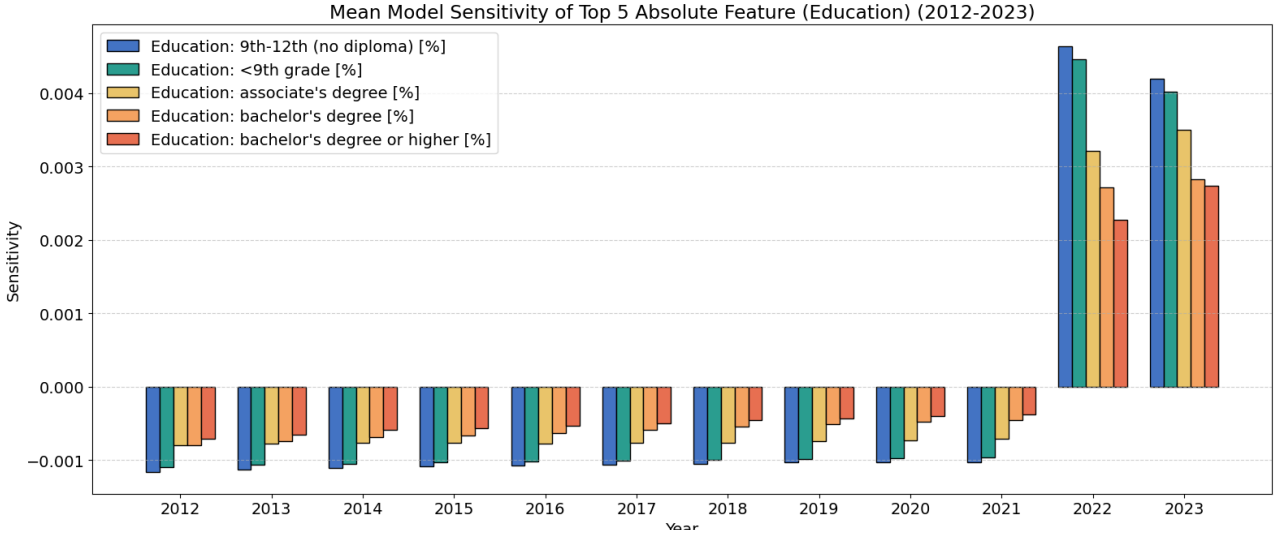

Realitity 'Median household income ($)' over continous County (2021)
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
160000
140000
120000
100000
80000
60000
40000
20000

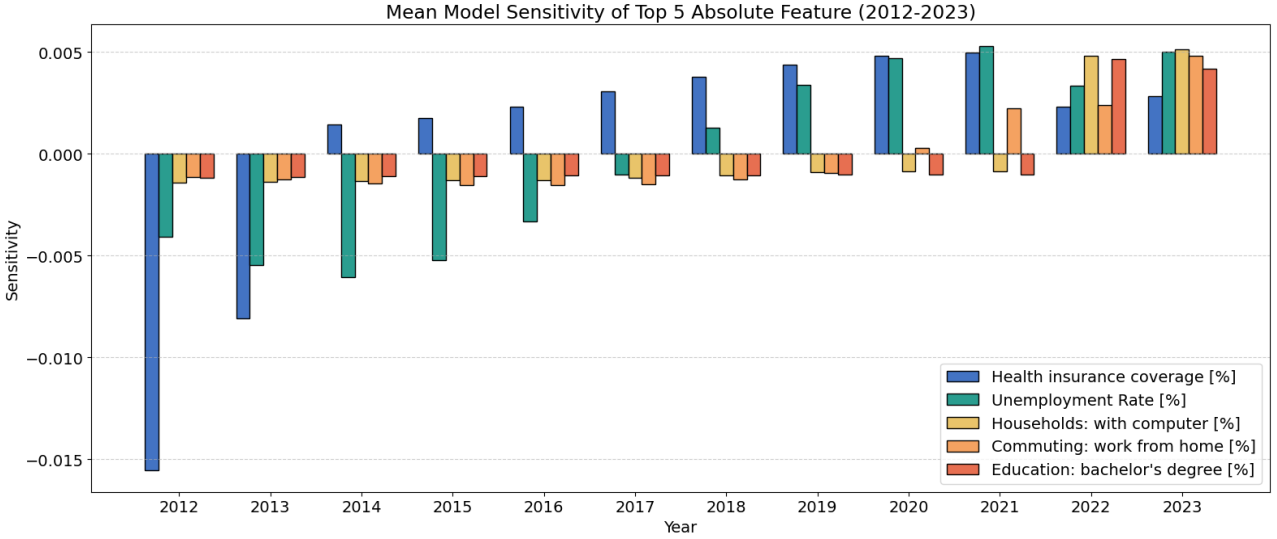

Mean Diffrence 'Median household income ($)' over Time (2012-2023)
Mean Diffrence in ($)
Year
3000
2500
2000
1500
1000
500
0
−500
2012
2014
2016
2018
2020
2022

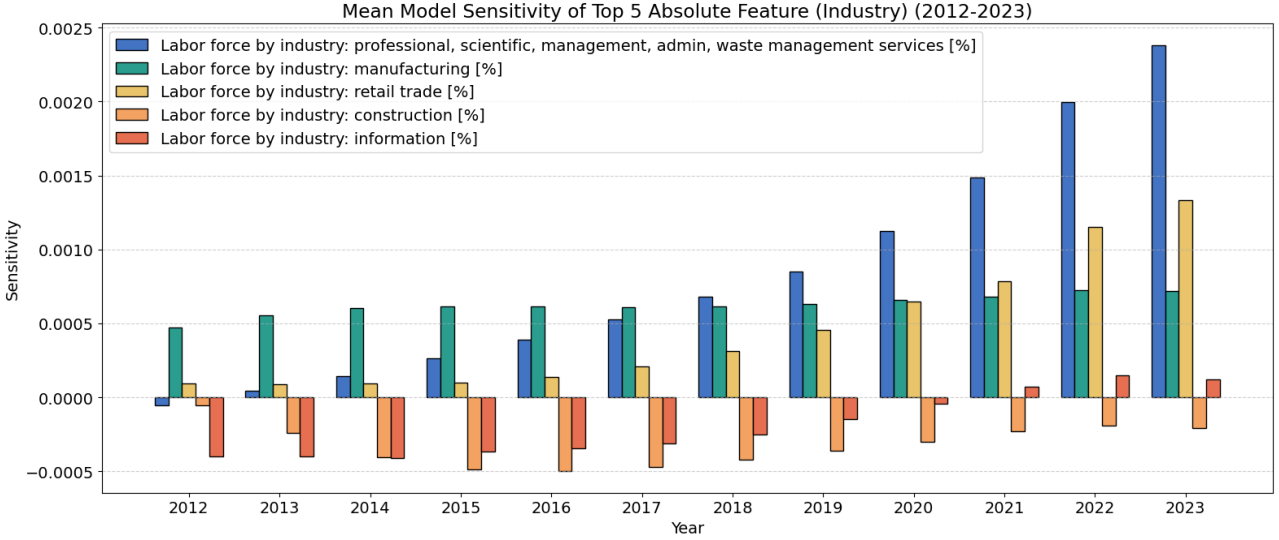

Prediction 'Median household income ($)' over continous County (2021)
Latitude
Longitude
50
45
40
35
30
25
−120
−110
−100
−90
−80
−70
160000
140000
120000
100000
80000
60000
40000
20000

Sensitivity Analysis









Germany (aggregate data)

Resources

DATASET

- Aggregated data: no individuals information
- County level: 390 / 476 Landeskreise
- Time span: 10 years
- Total samples: 3900

} limitation due to data quality

METHODS

Models

- Random Forest (RF) classification of income classes
- Multiple linear regression (MLR)

Train (80%) and Test (20 %) splits:

- Split randomly with 30 different seeds
- Mean and standard deviation estimation

Feature Importance:

- Random forest: Permutation feature importance (rank)
- MLR: Fitted coefficients

Data and shapefile sources:

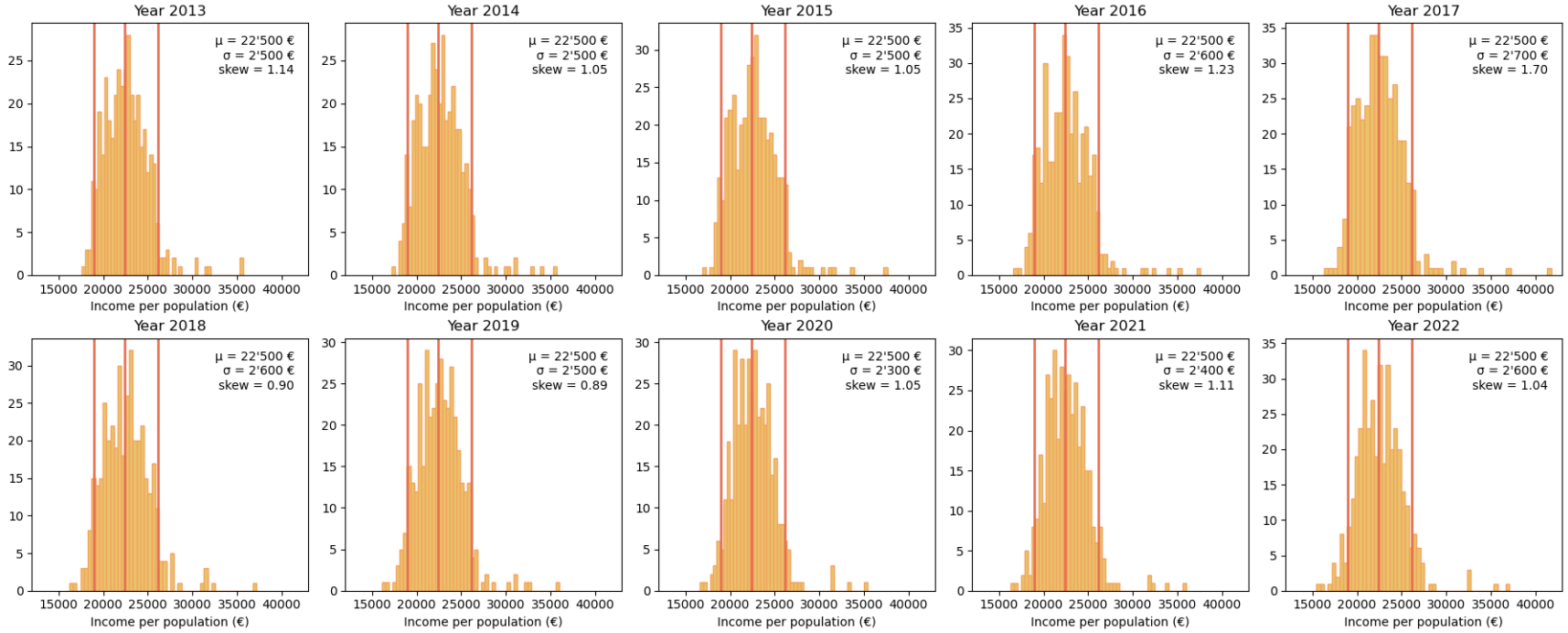
© Statistisches Bundesamt (Destatis), *GENESIS-Online*, https://www.regionalstatistik.de/genesis/online, accessed on: *[2025]*

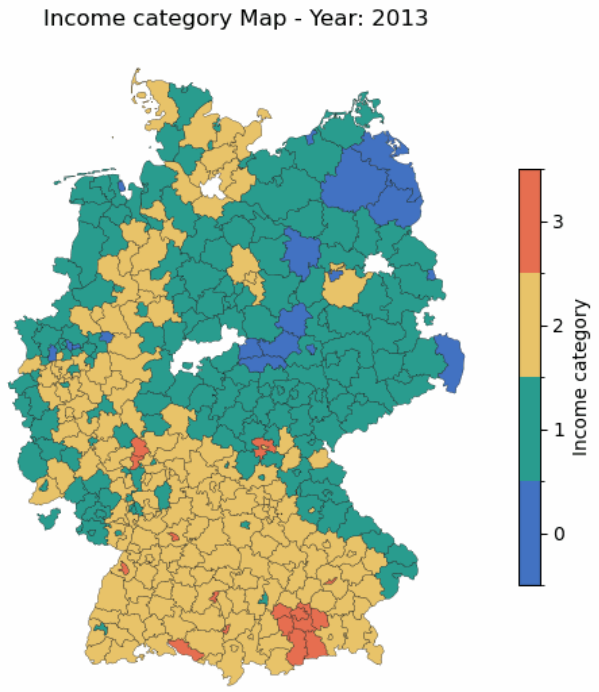
© **BKG** (2025) dl-de/by-2-0, Datenquellen:

https://sgx.geodatenzentrum.de/web_public/gdz/datenquellen/Datenquellen_vg_nuts.pdf

Income Categories

Category 0: < 5th percentile
Category 1: 5th - 50th percentile
Category 2: 50th - 95th percentile
Category 3: >95th percentile



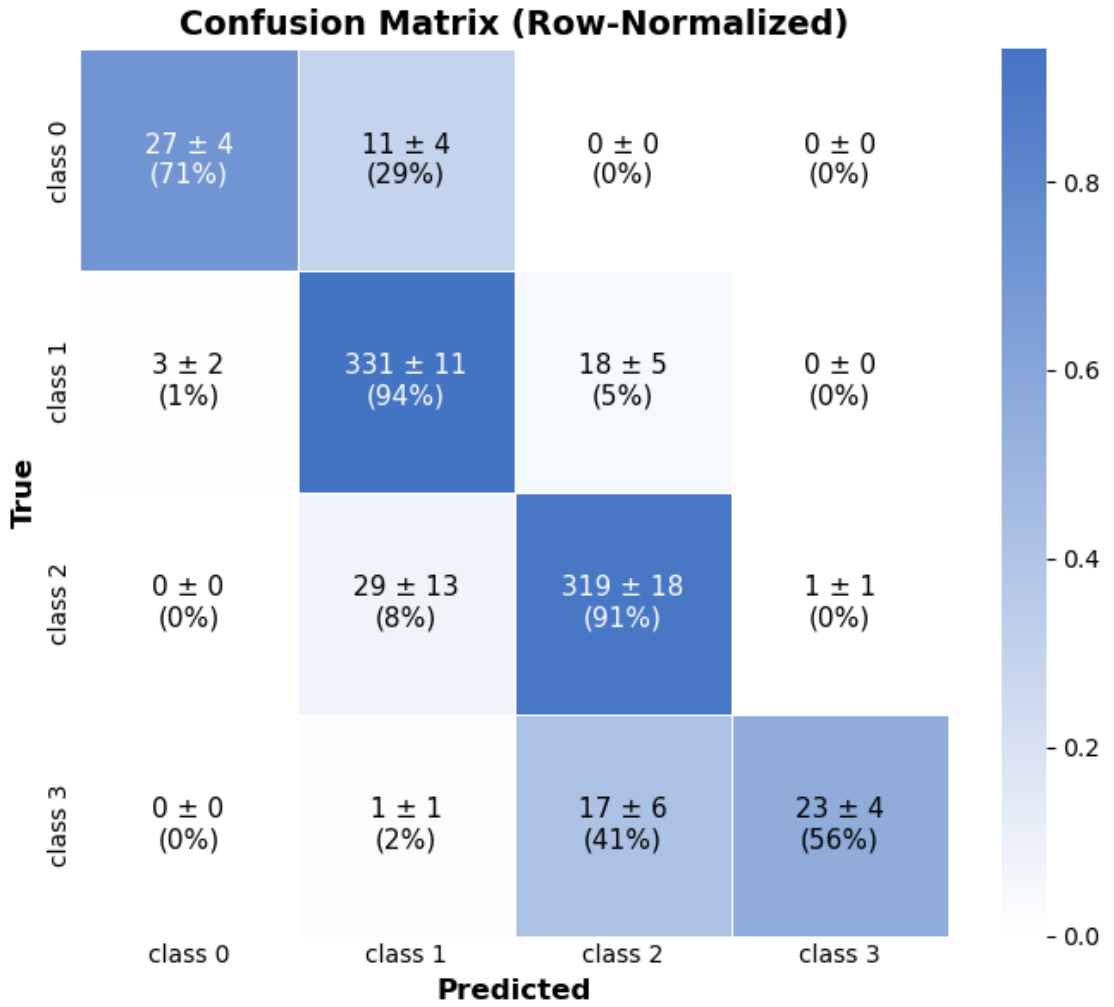


Classification

RANDOM FOREST

Overfitting

- Mean test accuracy: **0.897 ± 0.018**
- Mean train accuracy: 1 ± 0

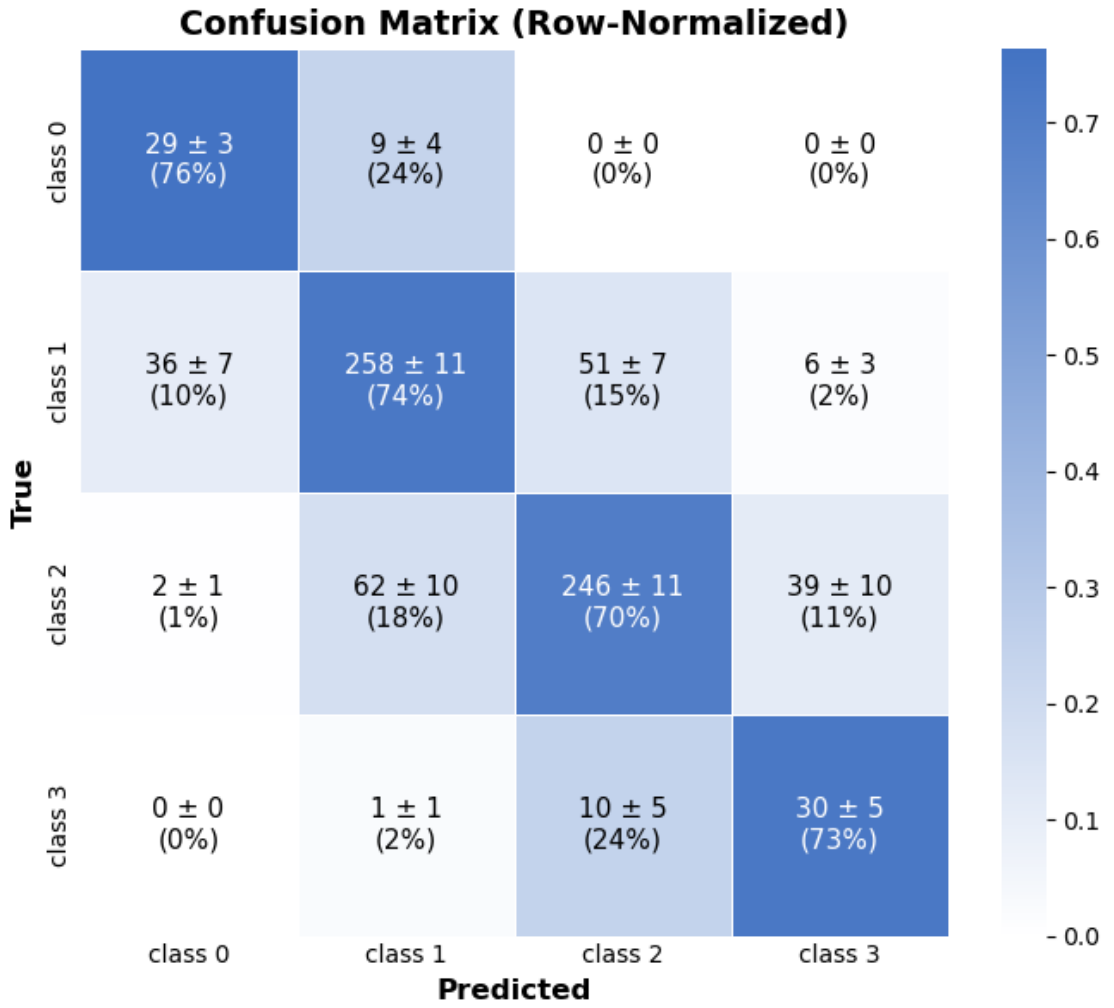


Classification

RANDOM FOREST

Preventing overfitting

- Mean test accuracy: **0.723 ± 0.019**
- Mean train accuracy: 0.780 ± 0.008



Classification

FEATURE IMPORTANCE

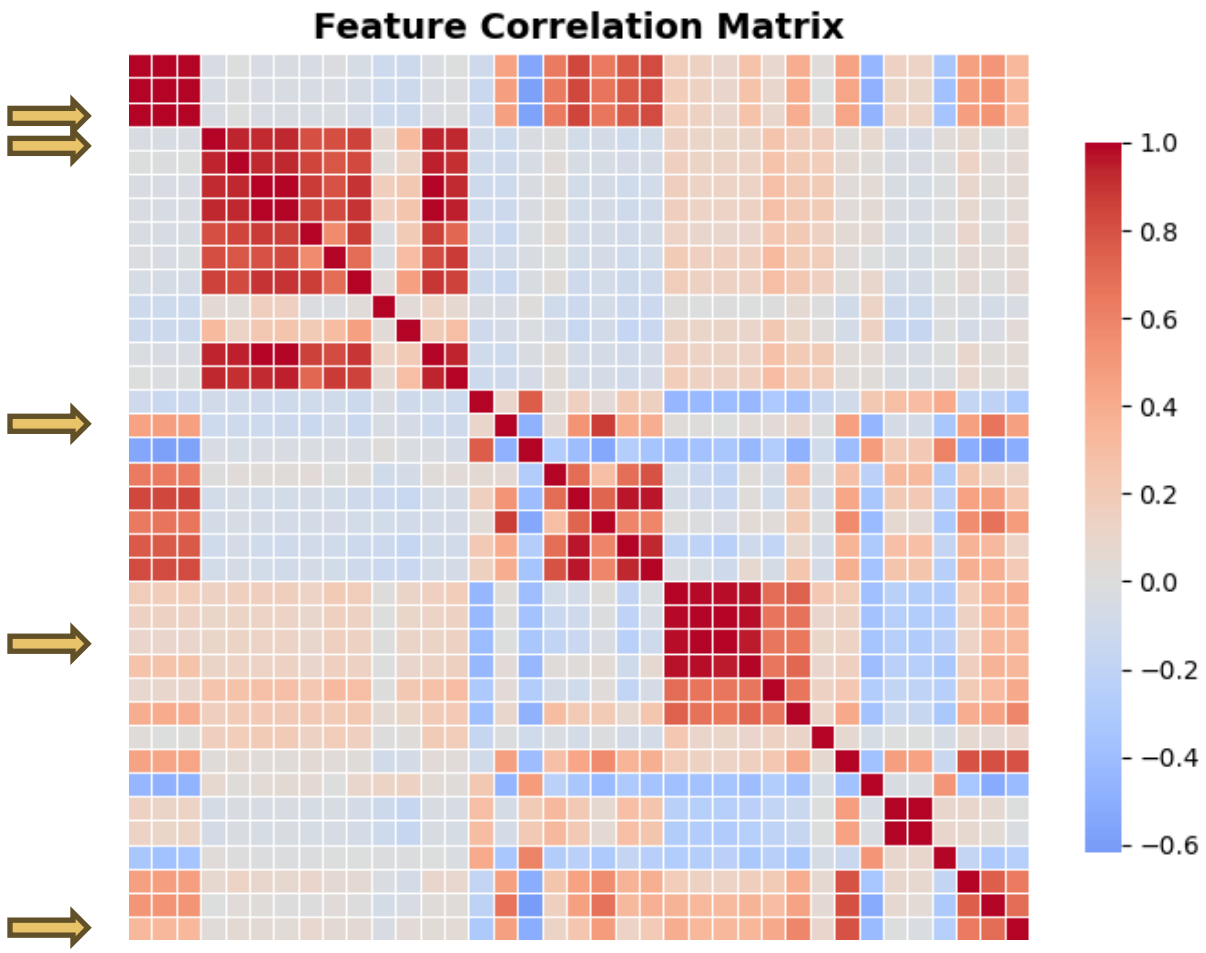
Overfitted model

Top 5 Features by average rank	Rank
Standard benefit (employable)	1.3 ± 0.7
Employed residents with academic degree (tot.)	2.6 ± 1.0
Total standard benefit	3.4 ± 1.8
Minimum income support benefits (tot.)	5.2 ± 2.5
Employed persons in Public & Social Services	5.9 ± 2.9

Not overfitted model

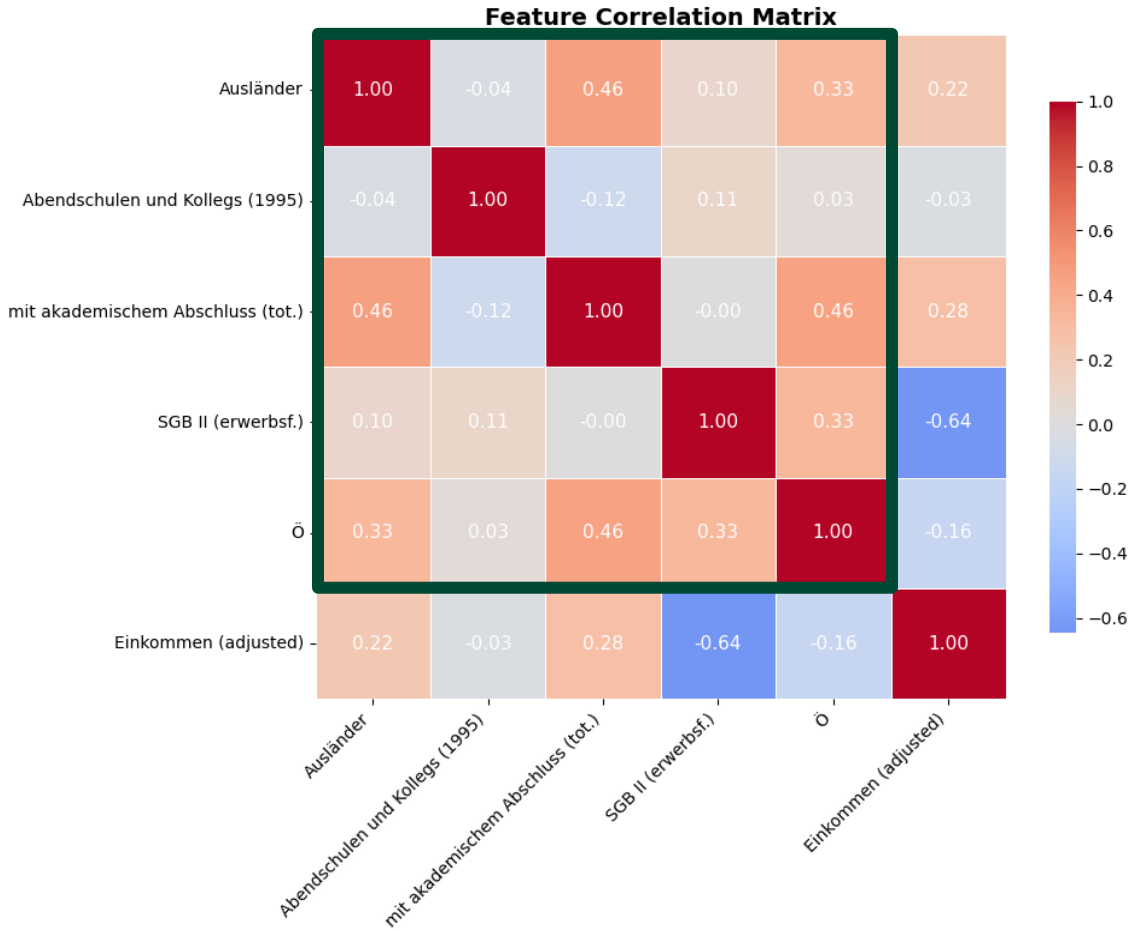
Top 5 Features by average rank	Rank
Sonderschulen/ Förderschulen (1995)	2.9 ± 1.5
Employed residents without vocational degree (tot.)	3.5 ± 3.3
Gymnasien (1995)	4.4 ± 2.8
Employed persons in Trade & Transport (G-J)	5.5 ± 4.1
Employed persons in Finance & Real Estate (V)	6.8 ± 4.5

Highly correlated features



- Highly correlated features hurt interpretability in MLR

Weakly correlated features



- **Subset of features with weak to moderate correlation**

Regression

MULTIPLE LINEAR REGRESSION

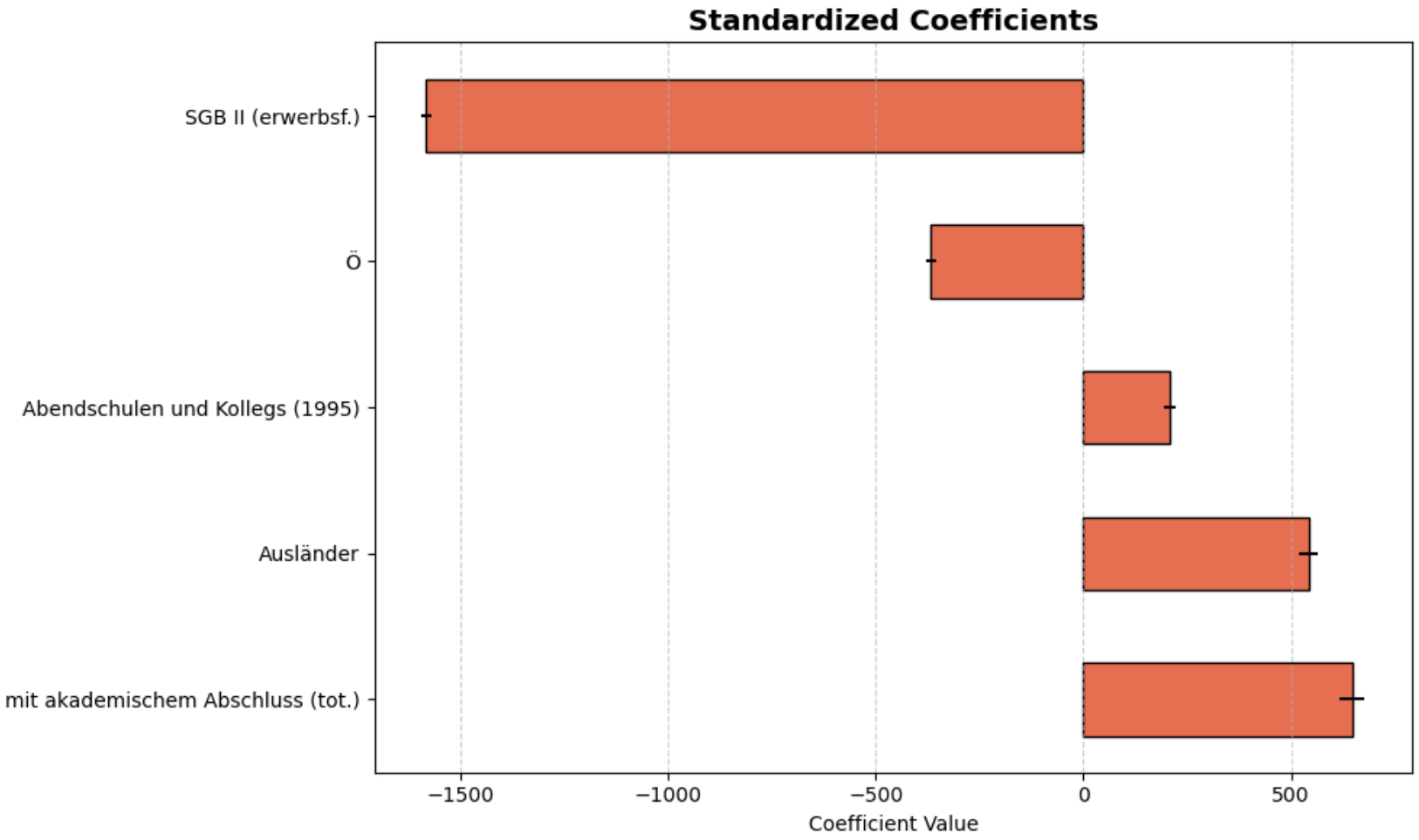
All features

- R^2 Score: **0.67 ± 0.02**
- R^2 Score (train): 0.680 ± 0.005
- RMSE: 1460 ± 80 €

Subset of features

- R^2 Score: **0.54 ± 0.02**
- R^2 Score (train): 0.547 ± 0.006
- RMSE: 1730 ± 90 €

Regression



Subset of features

- R^2 Score: 0.54 ± 0.02
- R^2 Score (train): 0.547 ± 0.006
- RMSE: 1730 ± 90 €

Conclusion

Education and social benefits are good predictors of income in regional aggregate data.

Number of schools per capita build in 1995 become good predictors for generalizable classifiers.

We could use a small ANN to allow non-linearities

USA (individuals data)

Source 6)
Total number of individuals: 15912393
Data from 2019-2023

Features:

AGE:	Age in years
SEX:	Biological sex
RACE:	Race codes
EDUC:	Education codes
UHRSWORK:	Usual hours worked per week
PWSTATE2:	Place of work: state
INCWAGE:	Income from wage
OCC_CATEGORY:	Occupation code

	YEAR	MULTYEAR	SAMPLE	SERIAL	CBSERIAL	HHWT	CLUSTER	STATEICP	STRATA	GQ	...
0	2023	2019	202303	1	2019010000088	2.0	2023000000013	41	260001	4	...
1	2023	2019	202303	2	2019010000096	14.0	2023000000023	41	70001	3	...
2	2023	2019	202303	3	2019010000153	4.0	2023000000033	41	80001	4	...
3	2023	2019	202303	4	2019010000198	17.0	2023000000043	41	80001	3	...
4	2023	2019	202303	5	2019010000205	11.0	2023000000053	41	280301	3	...
...	...	...	...	...	...	...	...	...	...	..	...
15912388	2023	2023	202303	7086486	2023001457972	15.0	2023070864863	68	30056	1	...
15912389	2023	2023	202303	7086486	2023001457972	15.0	2023070864863	68	30056	1	...
15912390	2023	2023	202303	7086487	2023001458196	15.0	2023070864873	68	20056	1	...
15912391	2023	2023	202303	7086488	2023001459187	7.0	2023070864883	68	10056	1	...
15912392	2023	2023	202303	7086488	2023001459187	7.0	2023070864883	68	10056	1	...

Sort by households
and identify
working child and working parent

=== Household 7027532 ===

	PERNUM	AGE	SEX	INCTOT	INCWAGE	working_parent	working_child
15781731	1	57	1	80000	80000	True	False
15781732	2	55	2	50000	50000	True	False
15781733	3	25	2	50000	50000	False	True

Parents' average income: 65000.00
Adult children's average income: 50000.00

=== Household 7027641 ===

	PERNUM	AGE	SEX	INCTOT	INCWAGE	working_parent	working_child
15781998	1	62	1	50400	50000	True	False
15781999	2	32	1	30200	30000	False	True

Parents' average income: 50400.00
Adult children's average income: 30000.00

=== Household 7027710 ===

	PERNUM	AGE	SEX	INCTOT	INCWAGE	working_parent	working_child
15782141	2	56	2	70000	70000	True	False
15782142	3	35	1	30000	30000	False	True
15782143	4	33	2	40000	40000	False	True

Parents' average income: 70000.00
Adult children's average income: 35000.00

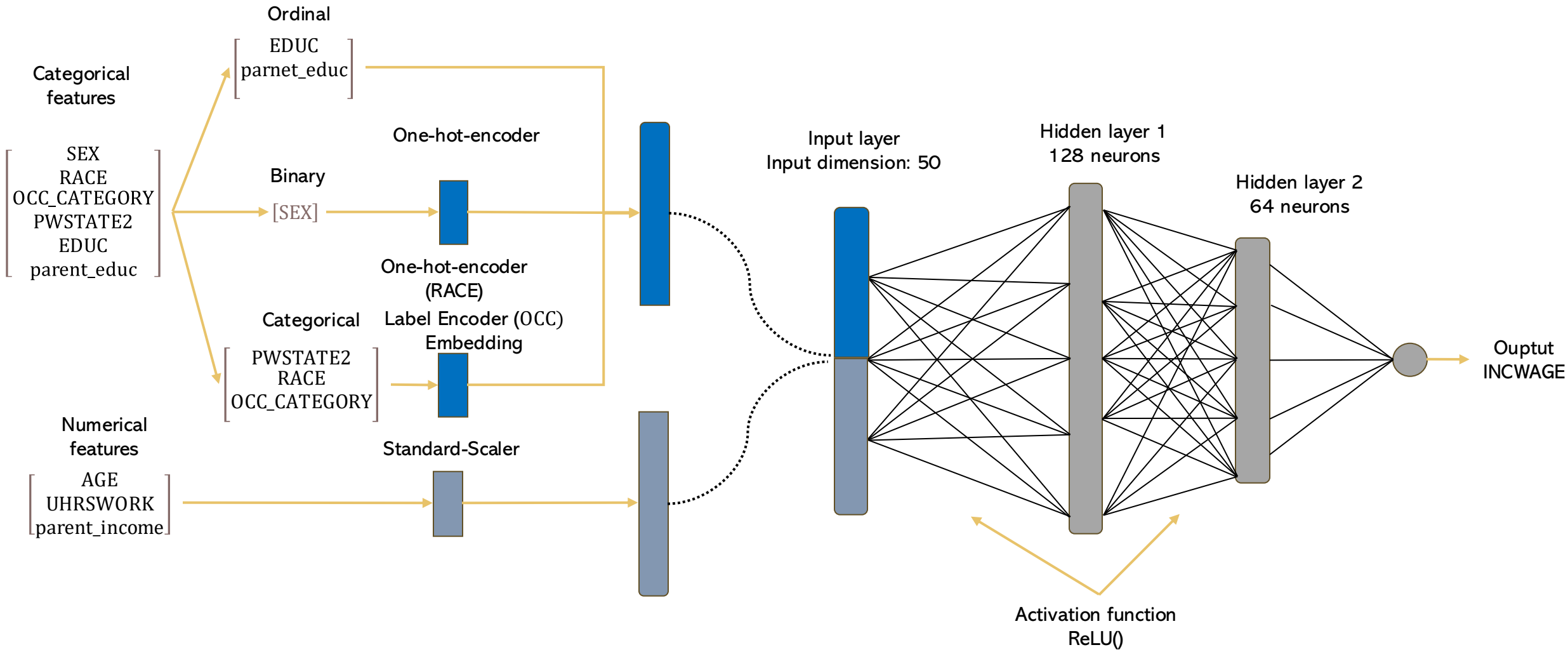
	AGE	SEX	RACE	EDUC	UHRSWORK	PWSTATE2	INCWAGE	parent_income	parent_educ	OCC_CATEGORY
0	25	1	1	6	43	1	29497	72799.0	7	Office and Administrative Support OCC
1	28	1	8	6	40	1	21238	63714.0	6	Installation, Maintenance, and Repair OCC
2	26	2	1	10	43	1	42358	55543.0	6	Education, Legal, Community Service, Arts, and...
3	29	1	1	10	40	1	100290	158105.0	10	Computer, Engineering, Science OCC
4	28	1	2	7	48	1	47195	27137.0	6	Production, Transportation, and Material Movin...
...	...	...	...	...	...	...	...	...	...	...
126903	25	1	1	6	40	56	28100	54000.0	11	Service OCC
126904	26	1	1	7	55	56	40000	50000.0	8	Installation, Maintenance, and Repair OCC
126906	27	2	1	10	40	56	34000	72350.0	7	Management, Buisness and Financial OCC
126907	26	2	1	11	40	31	64000	44500.0	10	Healthcare Practitioners and Technical OCC
126908	26	1	1	6	40	56	22000	43700.0	8	Service OCC

Total number of individuals: 118215

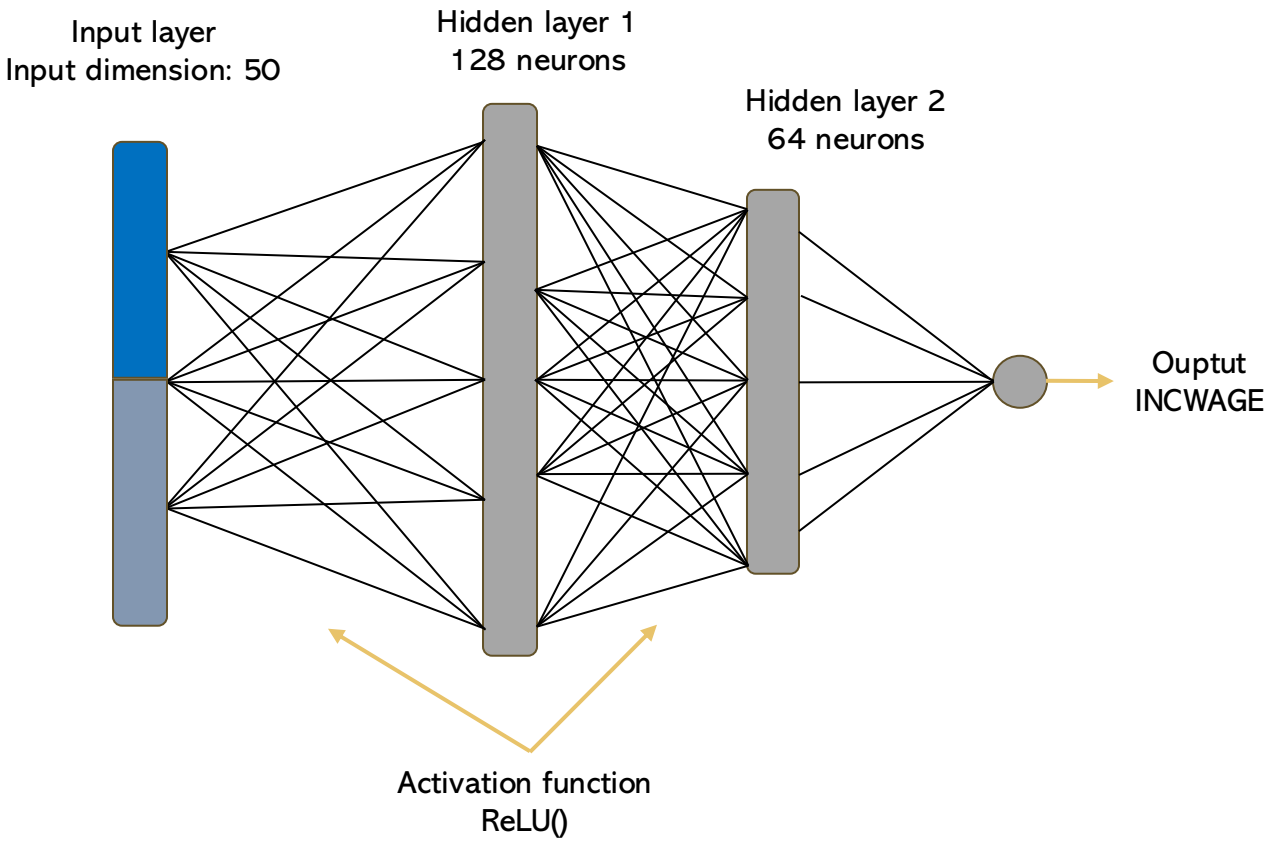
Working child:
- **EMPSTAT (employment status = 1 or 2 (self-employed))**
- **AGE: 25-35**
- **INCWAGE (Income from wage): 15000-300000**
- **UHRSWORK: >=35**

Working parent:
- **EMPSTAT (employment status = 1 or 2 (self-employed))**
- **INCTOT (Total income): 15000-300000**
- **UHRSWORK: >=35**

Model Architecture



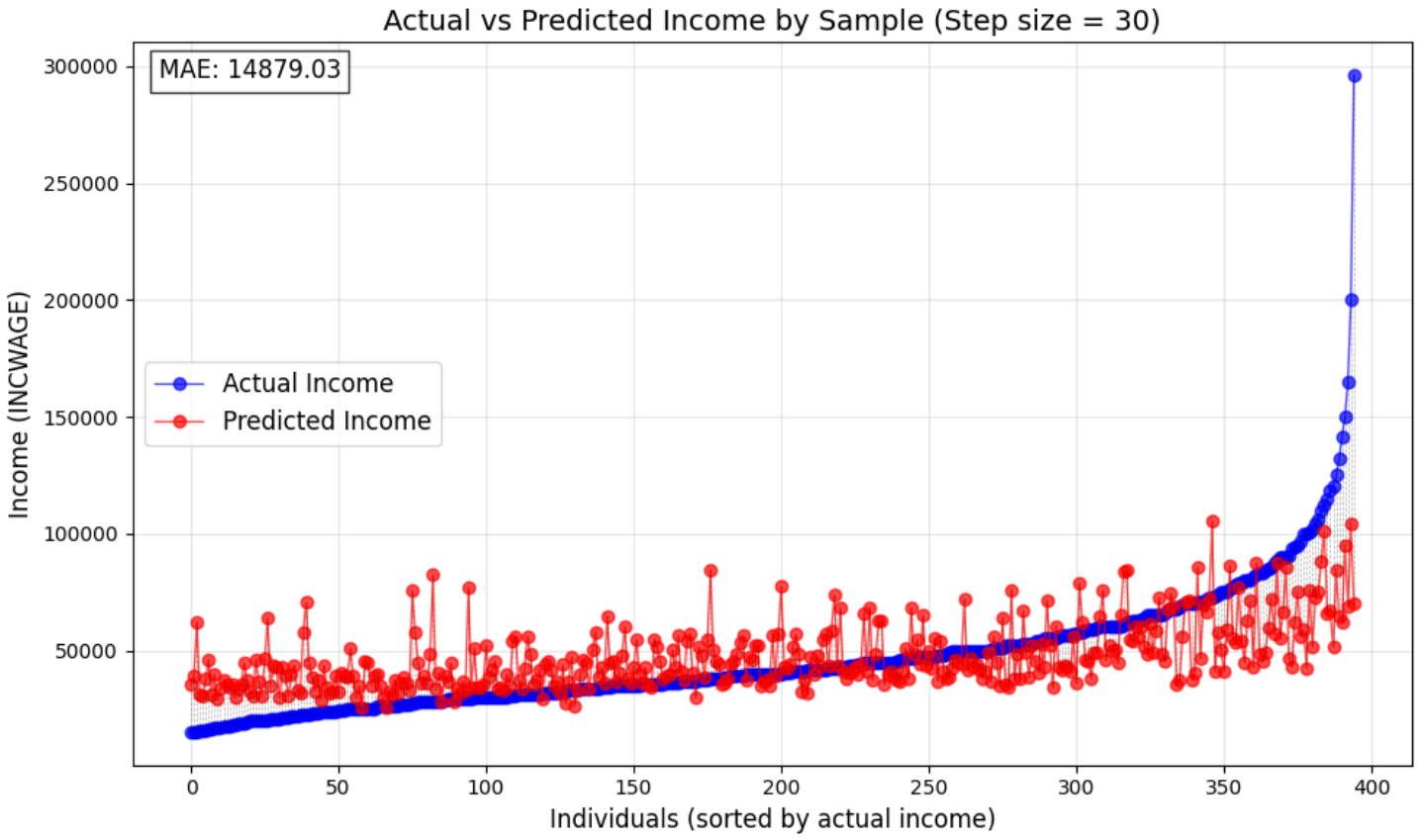
Model Architecture



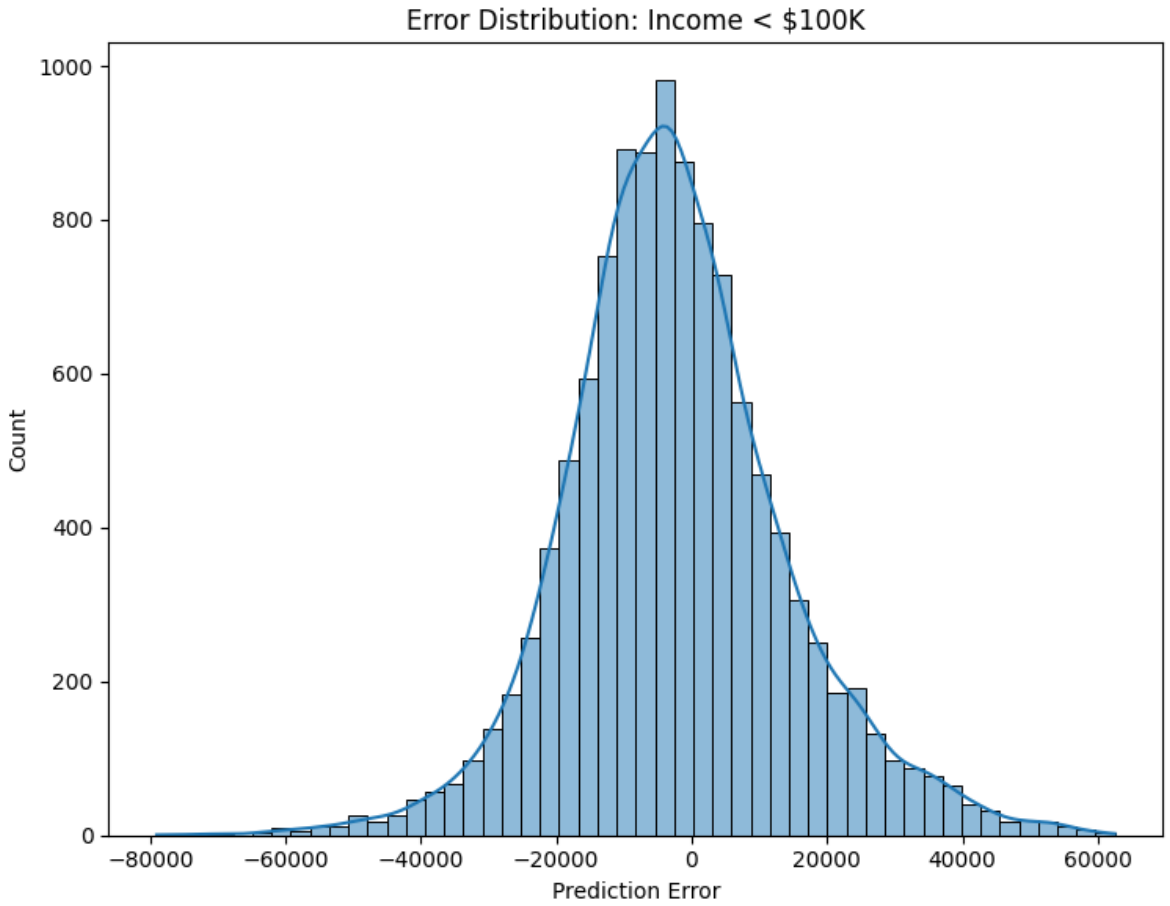
ML Model	
Optimizer	Adam
Loss	Mean Squared Error
Max Epochs	100
Batch sizes:	64
Training/Validation/Test	80%/10%/10%

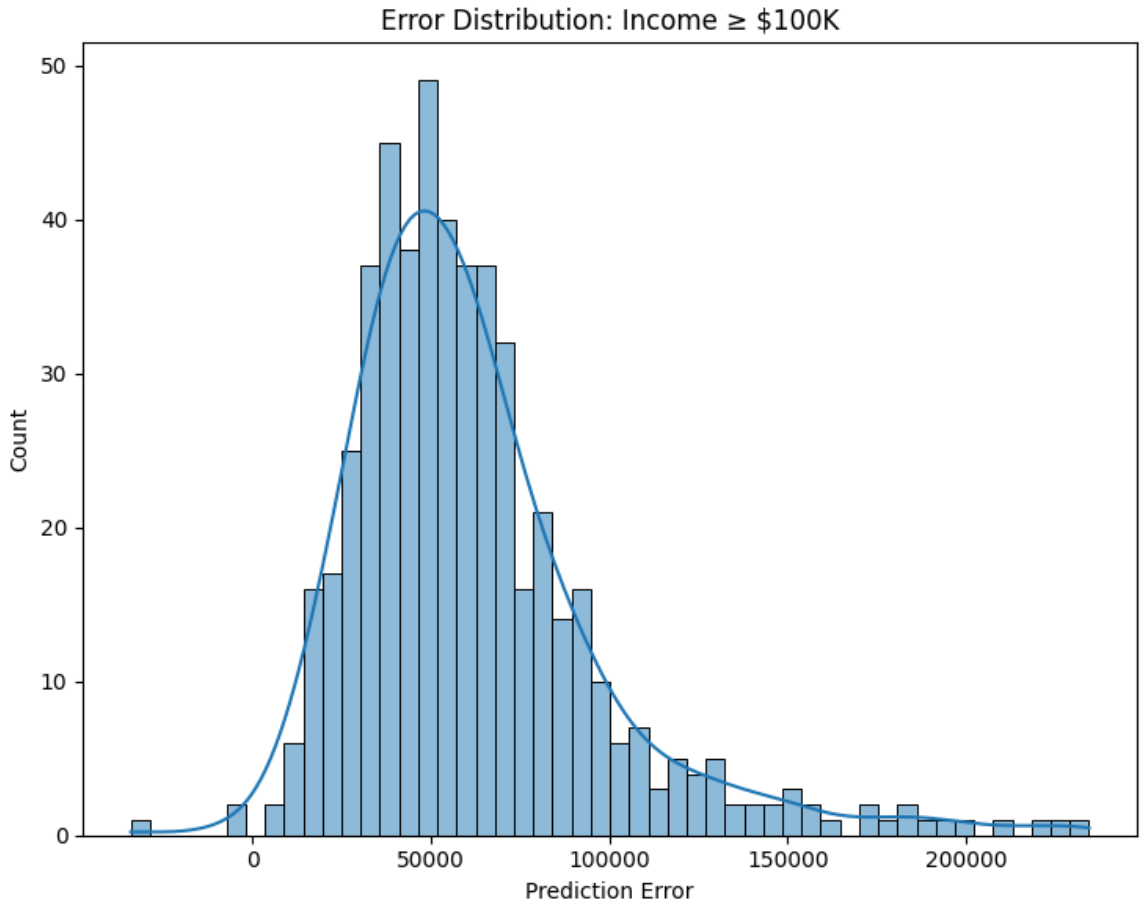
Accuracy

Metric	Loss
MAE	$14985
RMSE	$21975
R^2	0.29
Median Absolute Error	$10684
Mean Absolute Percentage Error	36.27

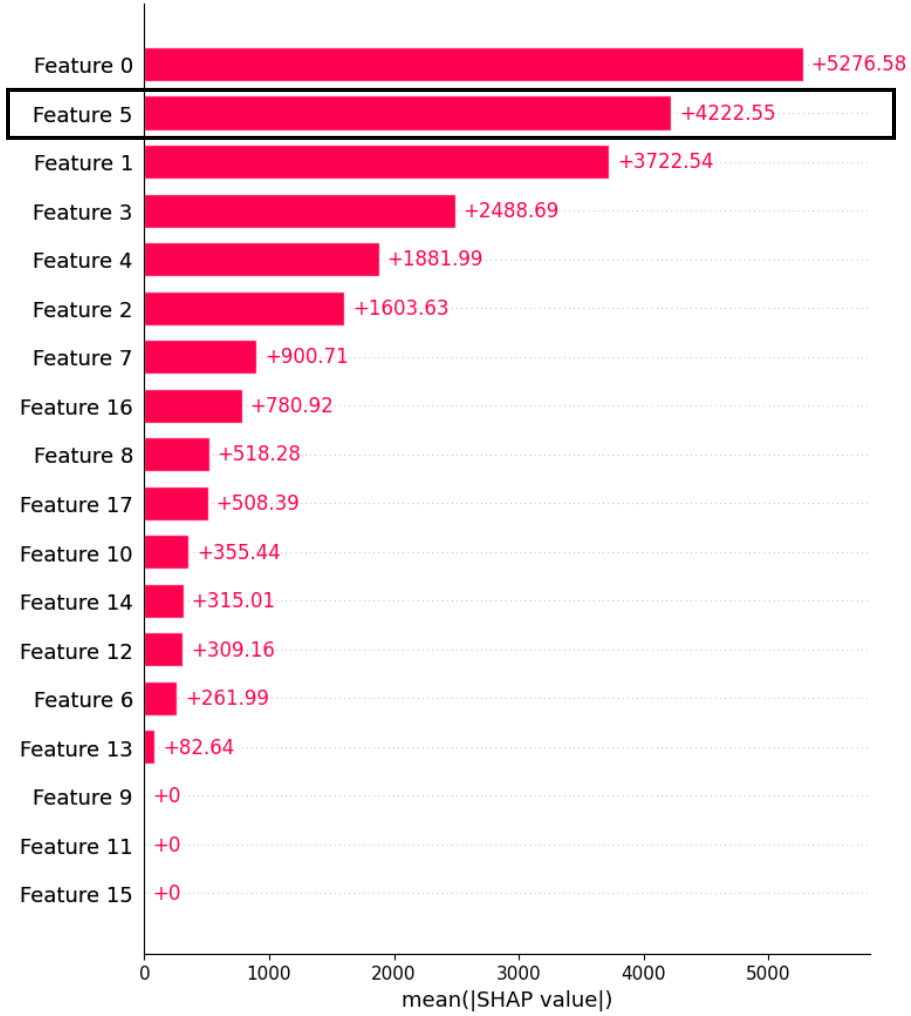


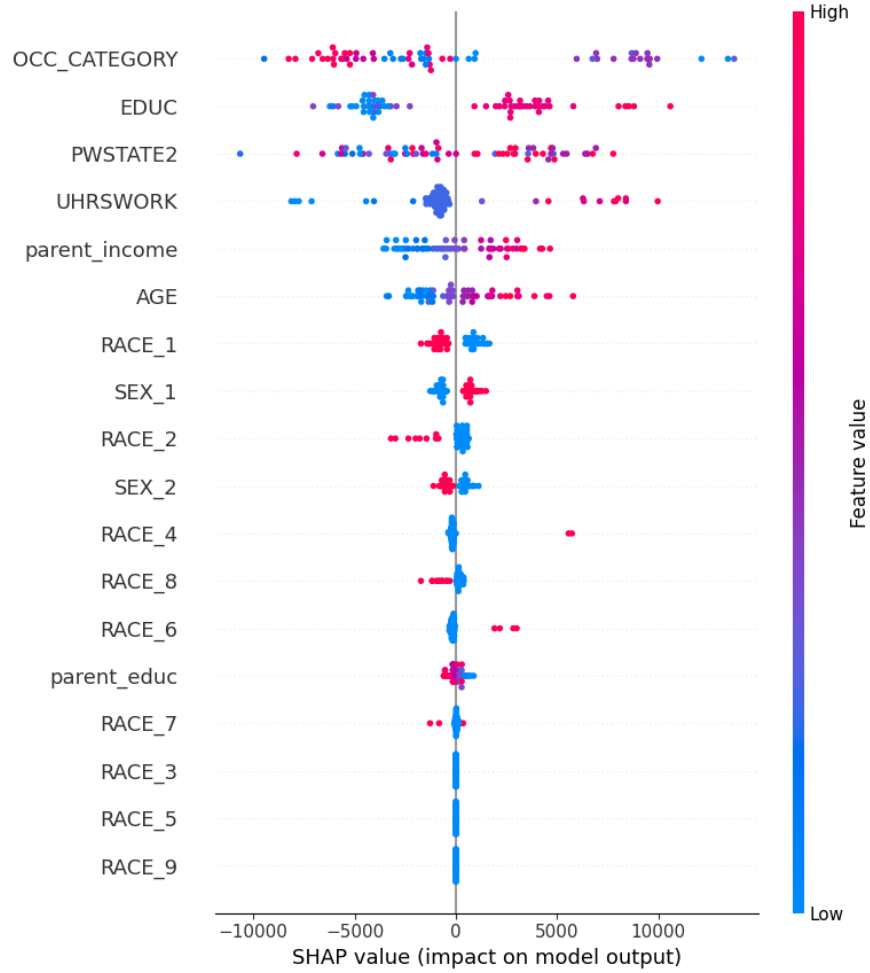
Accuracy


Error Distribution: Income < $100K
Count
Prediction Error
1000
800
600
400
200
0
−80000
−60000
−40000
−20000
0
20000
40000
60000


Error Distribution: Income ≥ $100K
Count
Prediction Error
50
40
30
20
10
0
0
50000
100000
150000
200000

Feature Analysis

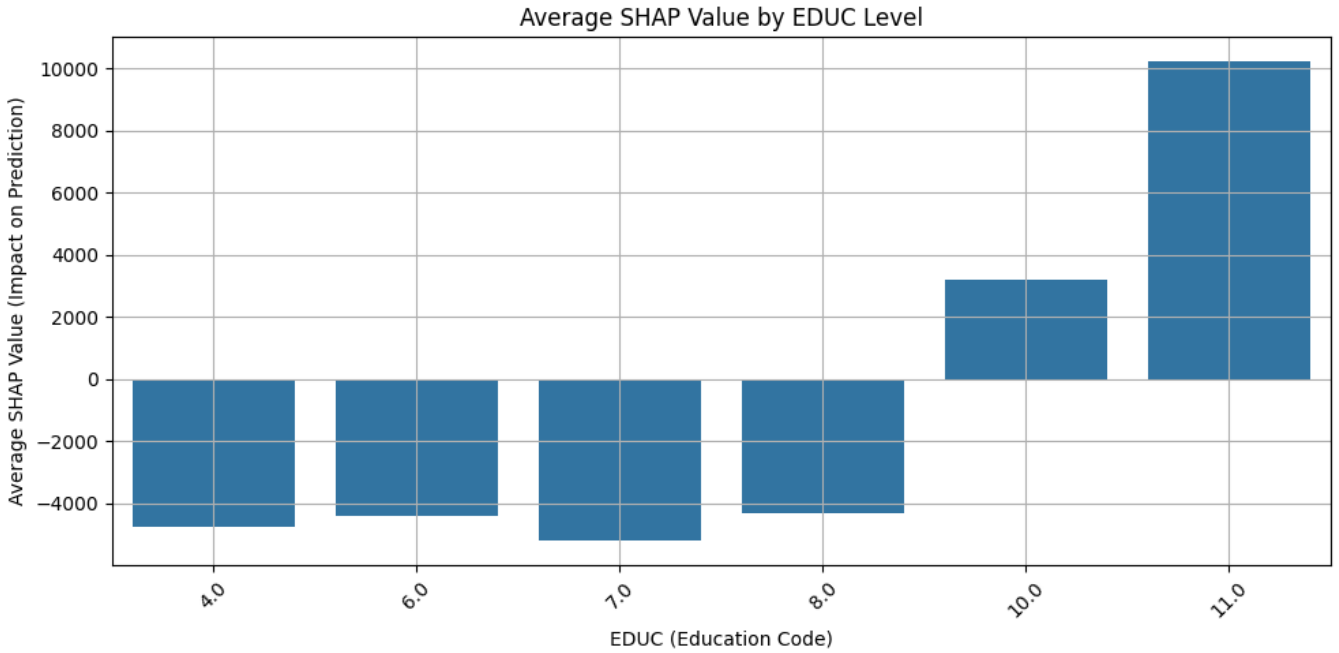


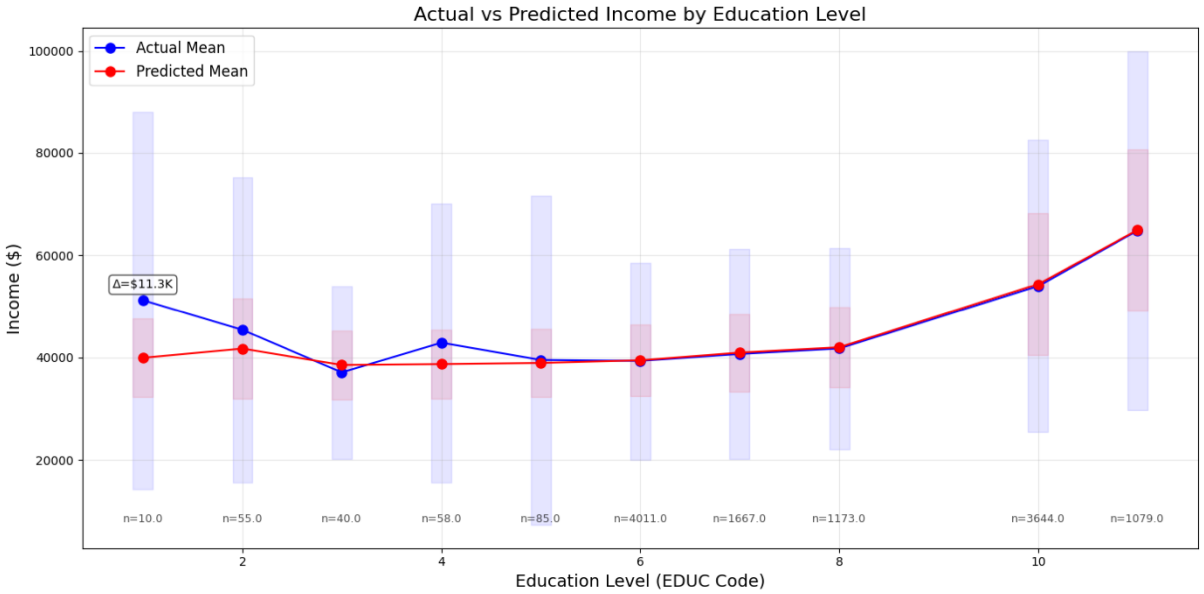


RACE = {
"White":1,
"Black":2,
"American Indian": 3,
"Chinese": 4,
"Japanese": 5,
"Other Asian": 6,
"Two major races": 8,
"Three or more major races": 9
}

SEX = {"Male":1,
"Female":2}

Feature Analysis





Conclusion

THOUGHTS

- **Education is one of the most important factors** indicating strong correlation between education and income
- **Higher Education leads to higher income**
- **Neural network can make moderately good predictions for income within a certain income range** (<100k)
- **Model struggles to predict incomes higher than 100k** (indicating that specific features or factors may be responsible for higher income levels)

IMPROVEMENTS

- **Utilize more features** (Feature gap mitigation)
- **Data sparsity --> more Data necessary**
- **Feature engineering** (Engineer features which target hidden determinats of high incomes)

India (household data)

India – Dataset IHDS

CHARACTERISTICS

- India Human Development Survey
- **Size**: 41’554 housesholds
- **Variaty**: 937 variables
- **Timespan**:
    1. **IHDS1**: 2004-2005
    2. IHDS2: 2011-2012
    3. IHDS3: ?

PREPROCESSING

- **Pick** the interesting features
- **Cleaning** the data in terms of
    1. Positive Income
    2. Missing values: ‘’ or ‘ ‘
    3. Numeric Data Type
- **Convert** to Tensors and Dictionaries

Features

Location
- Stateid, Distid, Distname

Education
- Highest male / female education
- At least one literate adult

Family
- Number of children in HH
- Number of persons in HH

Geographic
- Urban
- Metro

Religion and Caste
- Brahmins, …, Dalits
- Muslims, Sikhs, Jains, Christians

Household assests
- Computer, Car, HHassets

Features

Location
- Stateid, Distid, Distname

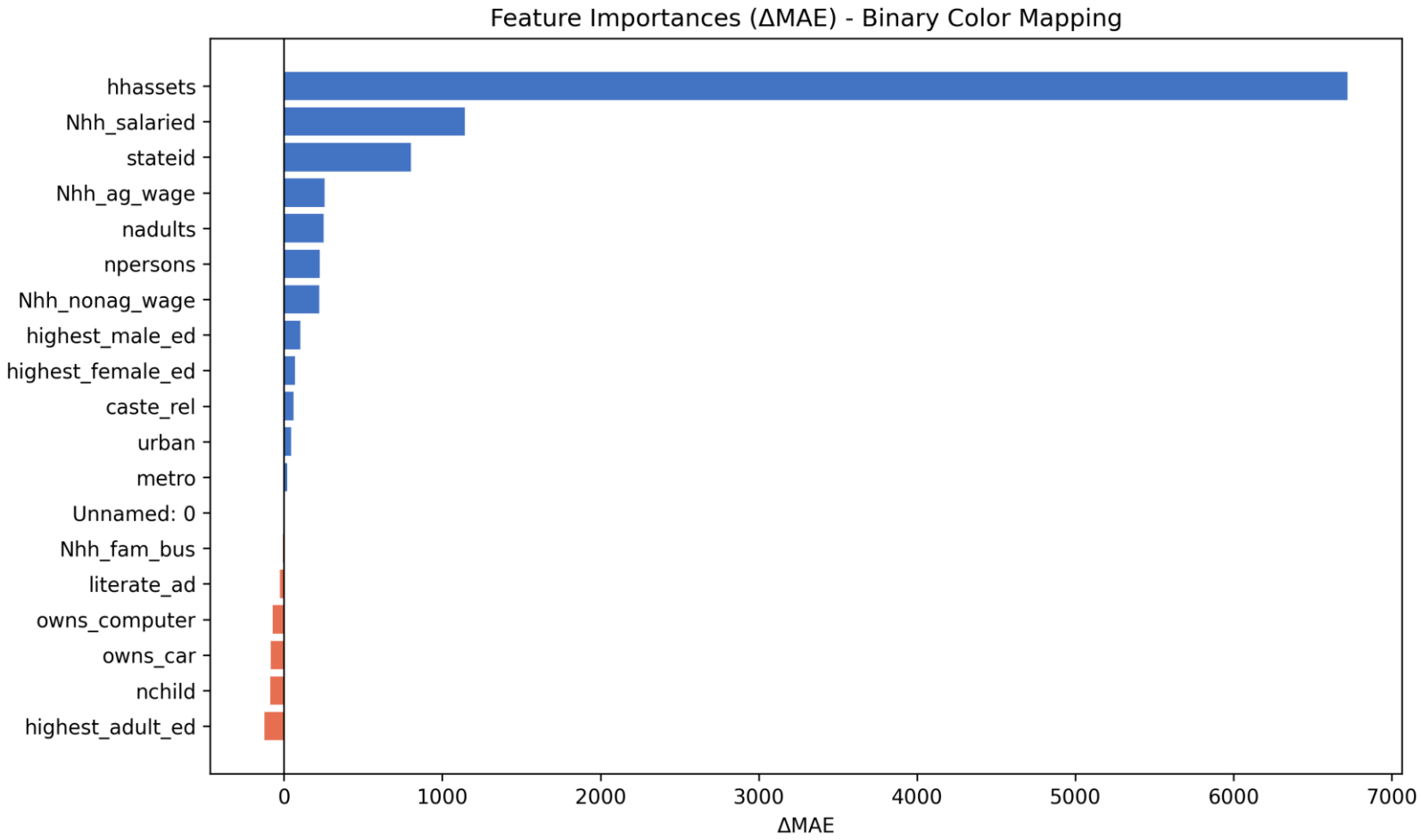
Education
- Highest male / female education
- At least one literate adult

Family
- Number of children in HH
- Number of persons in HH

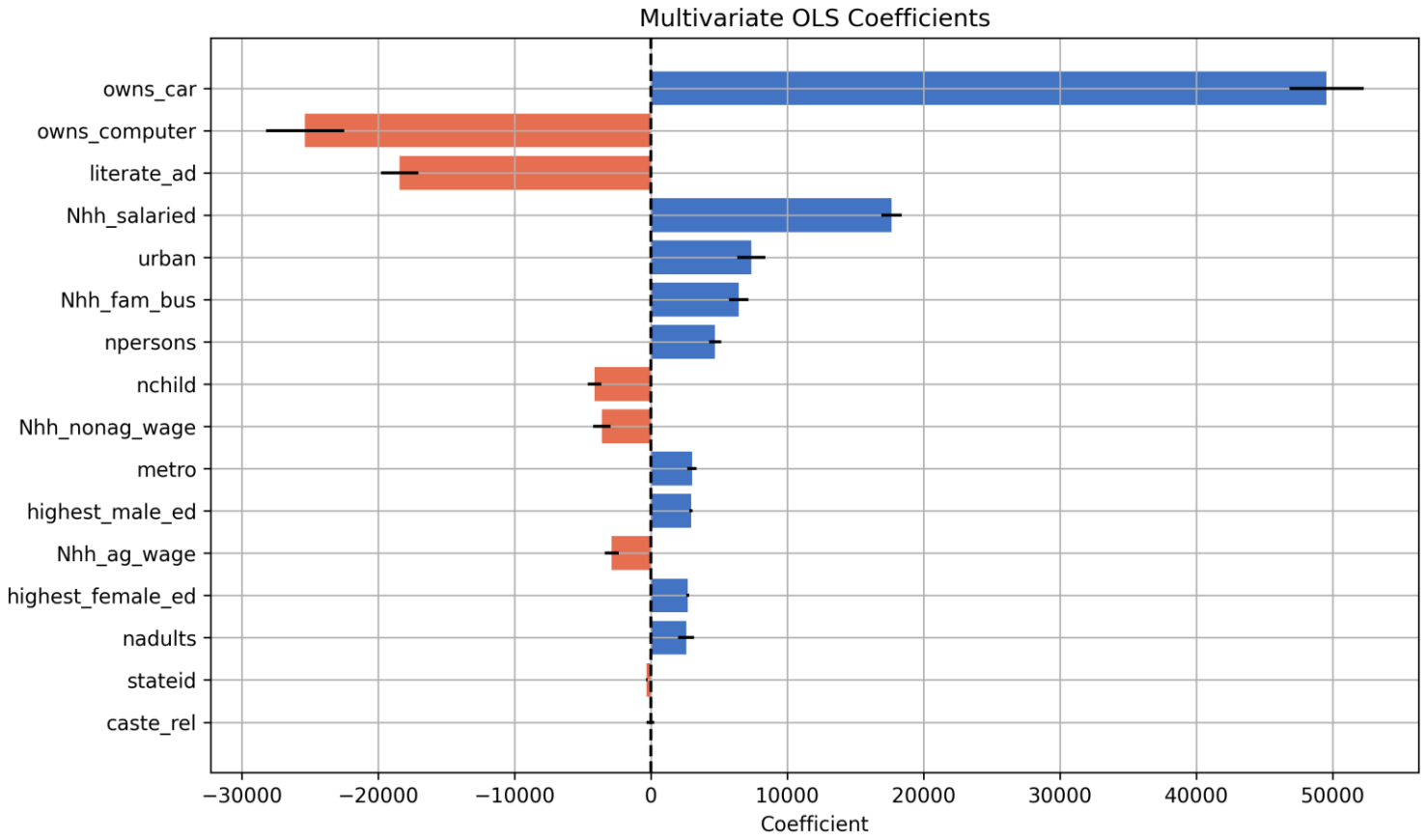
Geographic
- Urban
- Metro

Religion and Caste
- Brahmins, …, Dalits
- Muslims, Sikhs, Jains, Christians

Household assests
- Computer, Car, **HHassets**



Regression

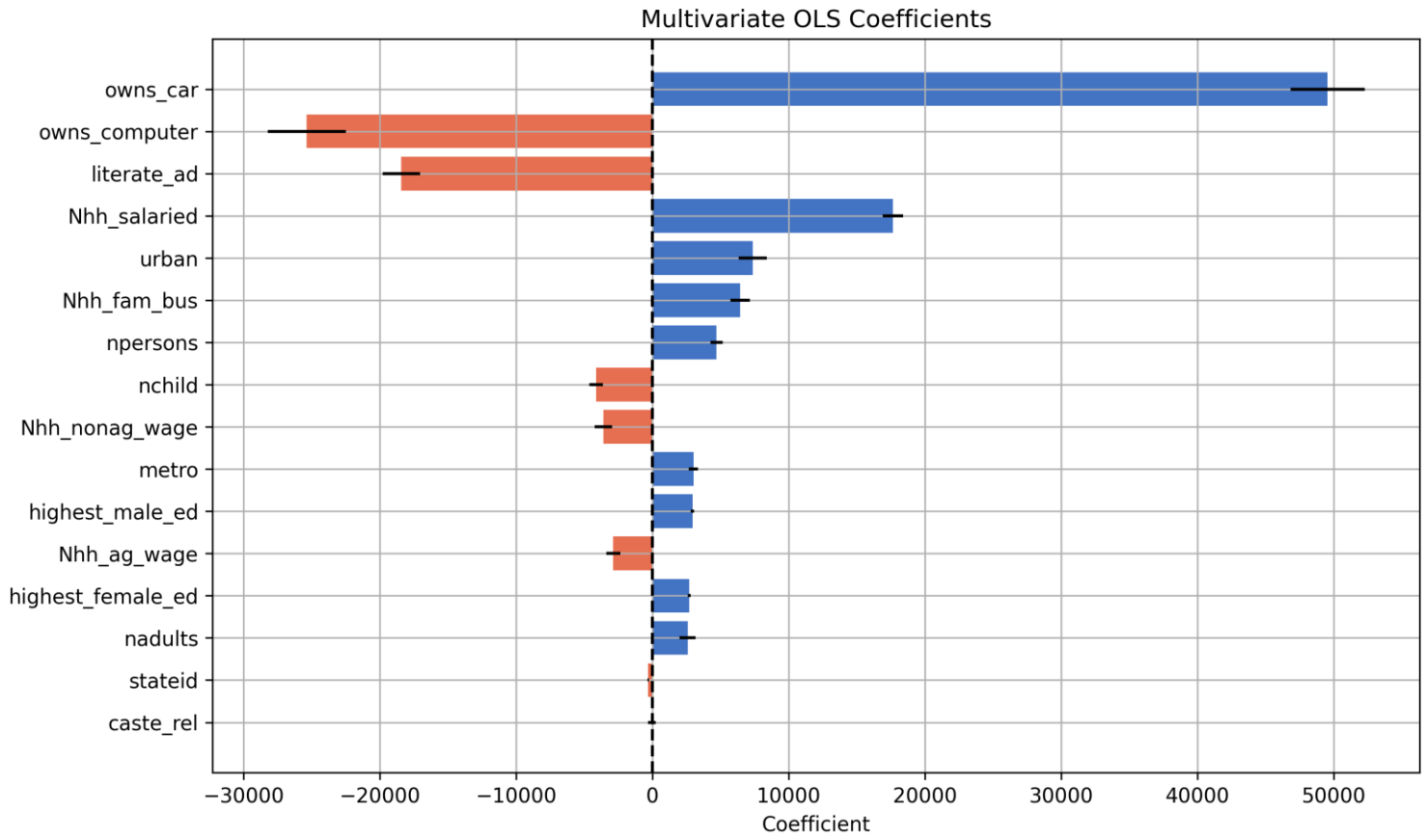

Multivariate OLS Coefficients
owns_car
owns_computer
literate_ad
Nhh_salaried
urban
Nhh_fam_bus
npersons
nchild
Nhh_nonag_wage
metro
highest_male_ed
Nhh_ag_wage
highest_female_ed
nadults
stateid
caste_rel
−30000
−20000
−10000
0
10000
20000
30000
40000
50000
Coefficient

Regression

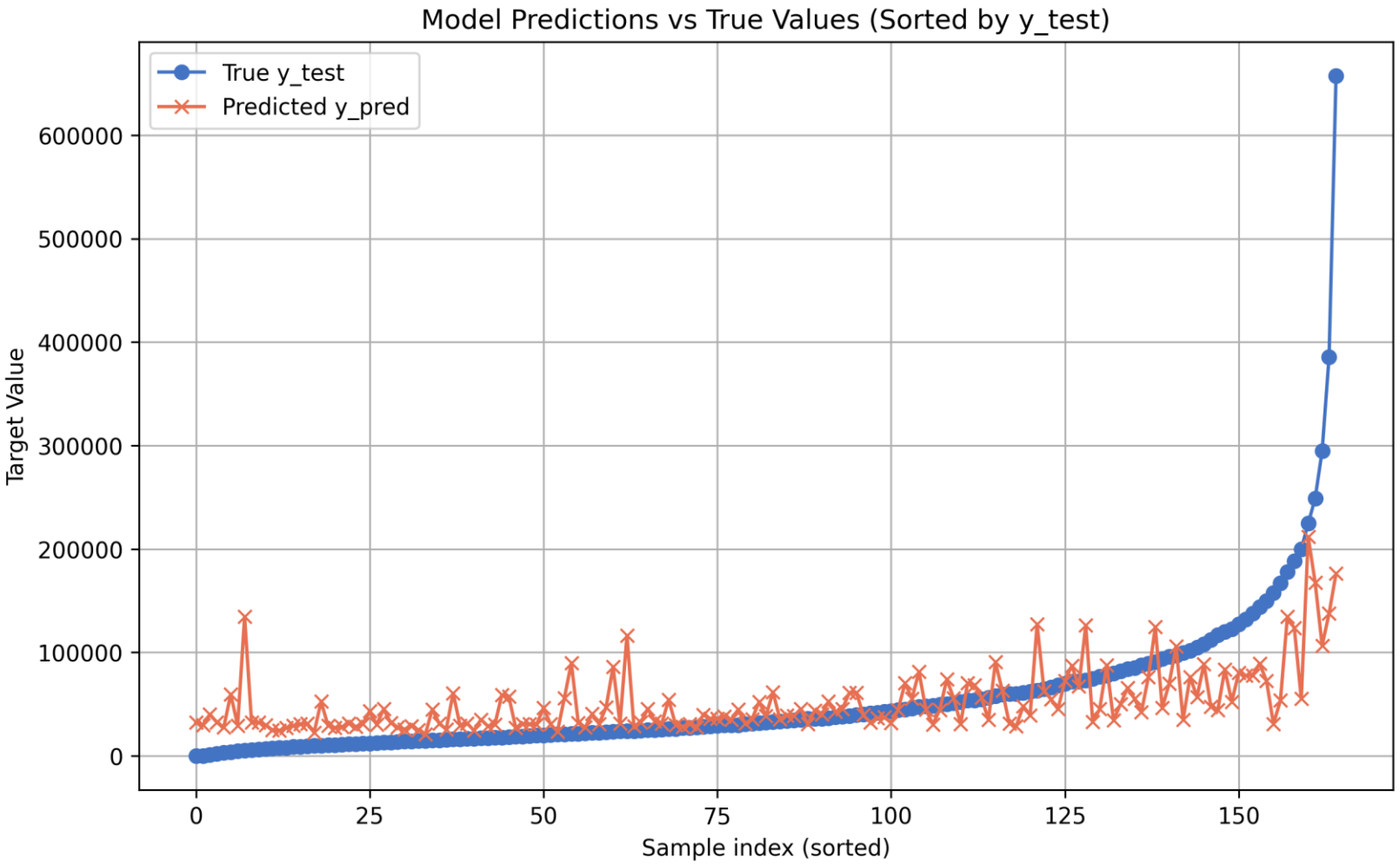
Metric	Loss
R2 Score	0.206
Skewness	5.6
Kurtosis	61

High coefficients

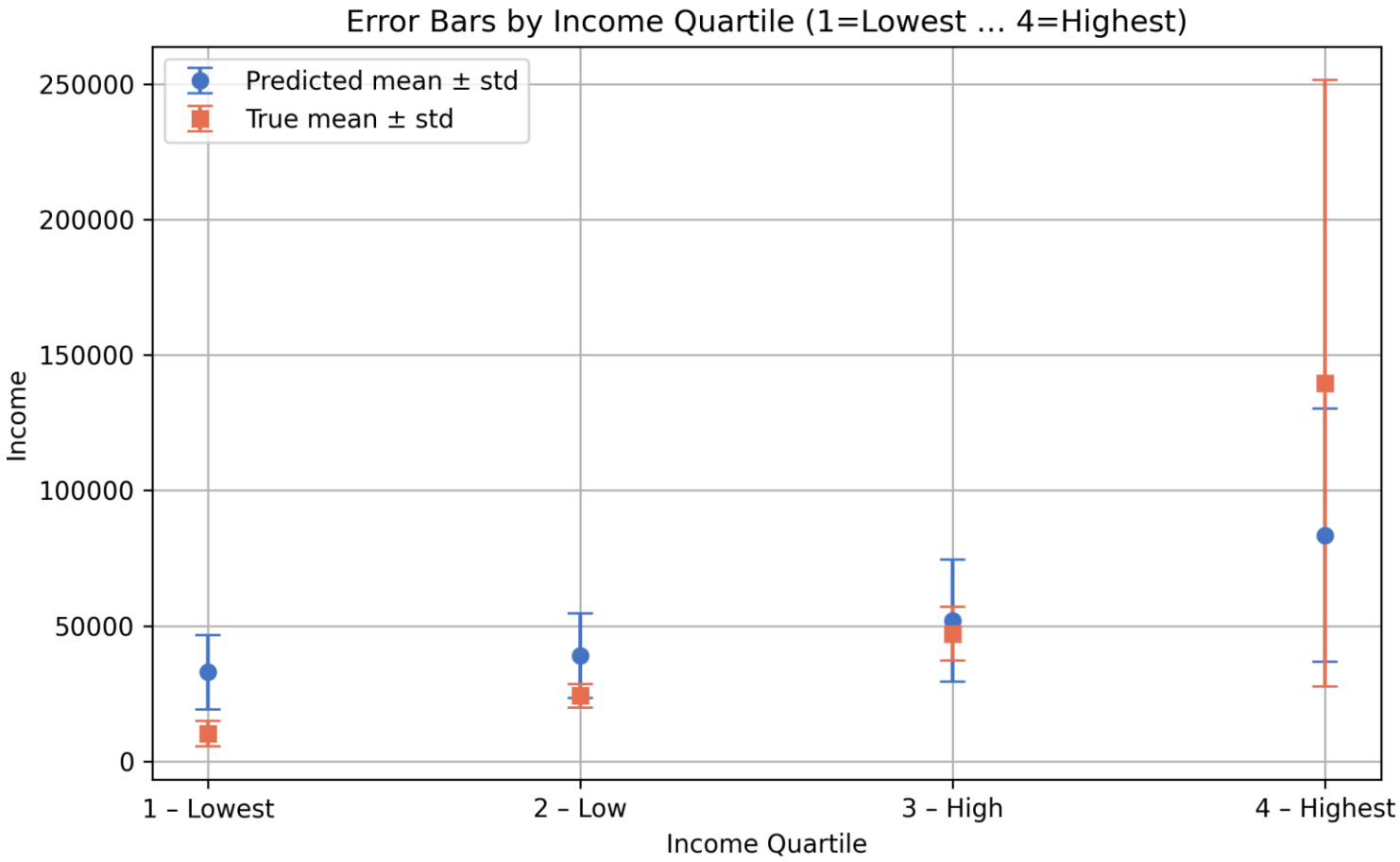
Unreasonable coefficients



Predictions



Accuracy

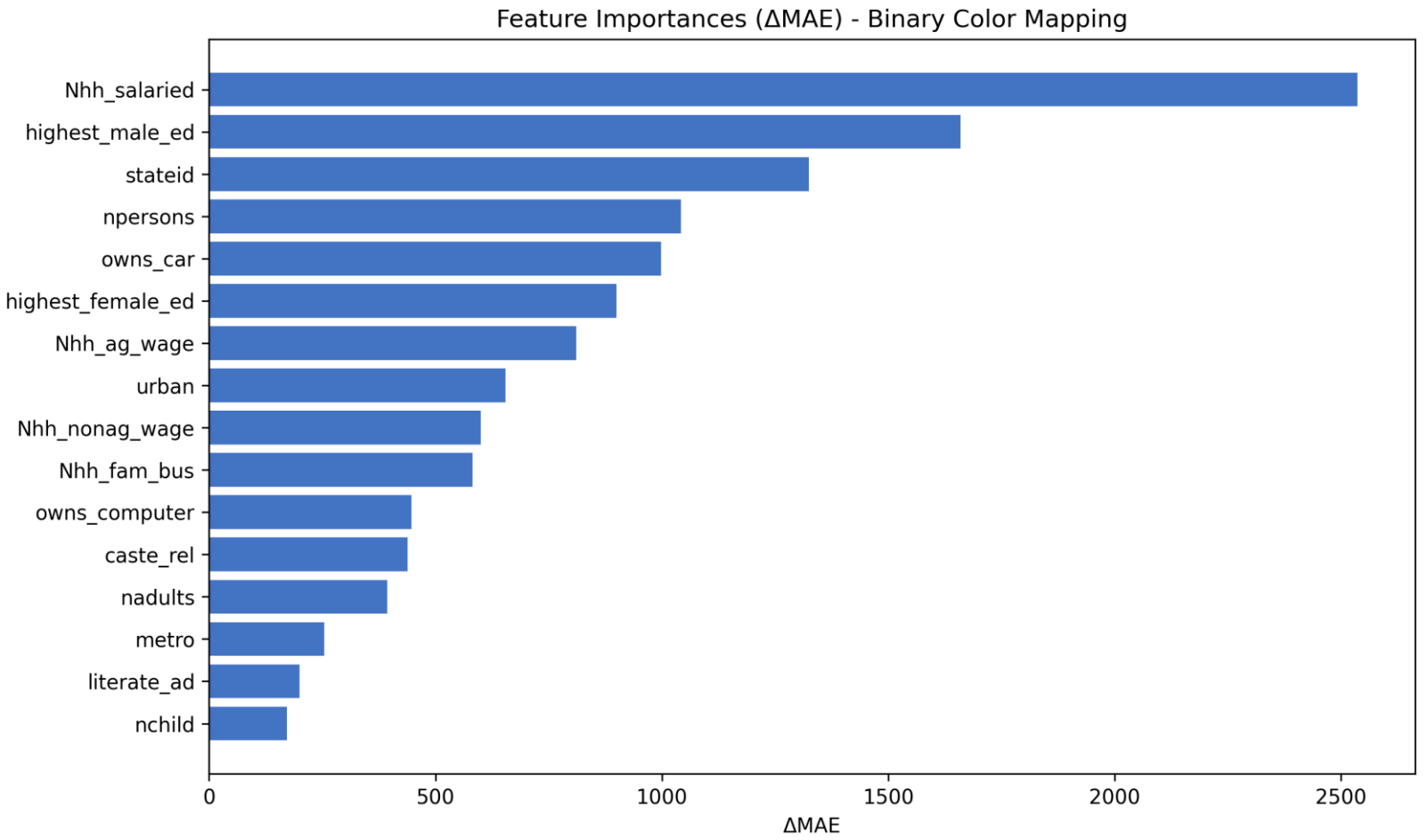

Error Bars by Income Quartile (1=Lowest ... 4=Highest)
Predicted mean ± std
True mean ± std
Income
250000
200000
150000
100000
50000
0
1 – Lowest
2 – Low
3 – High
4 – Highest
Income Quartile

Accuracy (cont.)

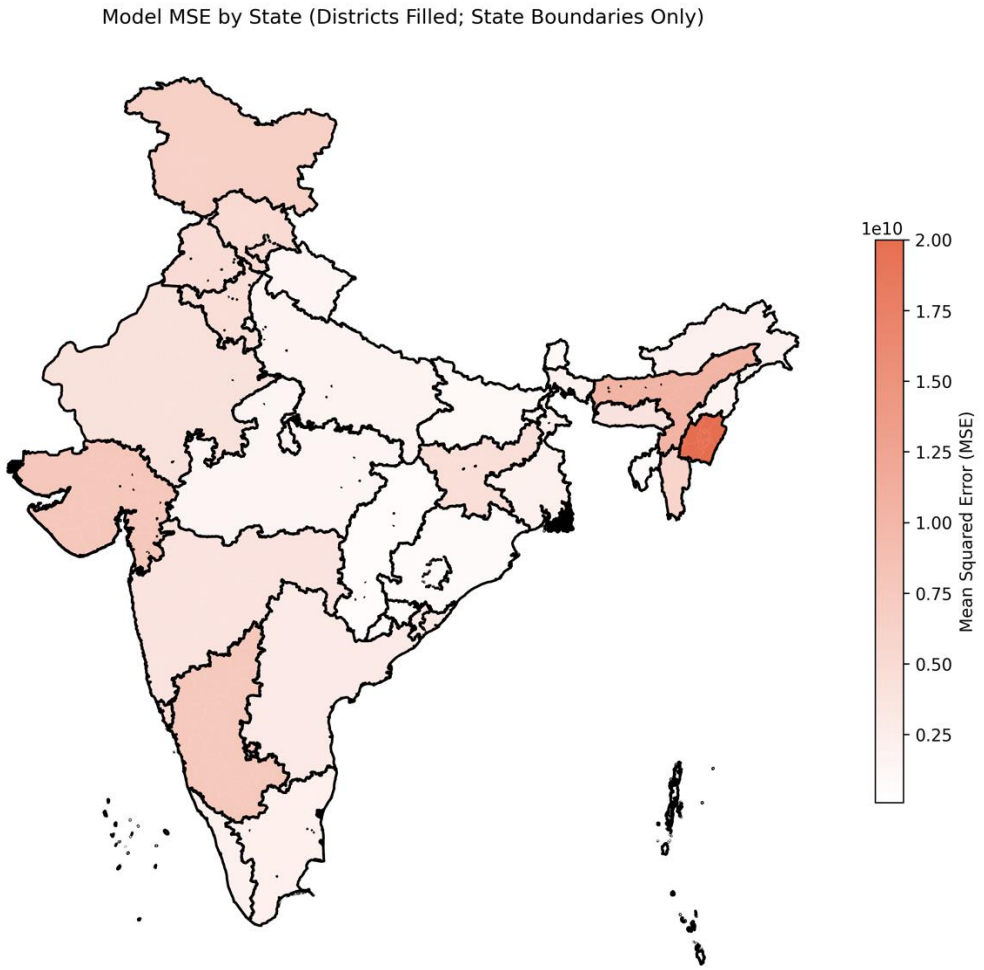
R2 Score	Skewness	Kurtosis
0.309	2	26

Metric	Loss	Conversion	Loss
L1 Loss **MAE**	29’397 Rupee	0,0098	287 CHF
1 – Lowest	22’659 Rupee	“	222 CHF
2 – Low	15’028 Rupee	“	147 CHF
3 – High	15’685 Rupee	“	153 CHF
4 – Highest	64’234 Rupee	“	629 CHF

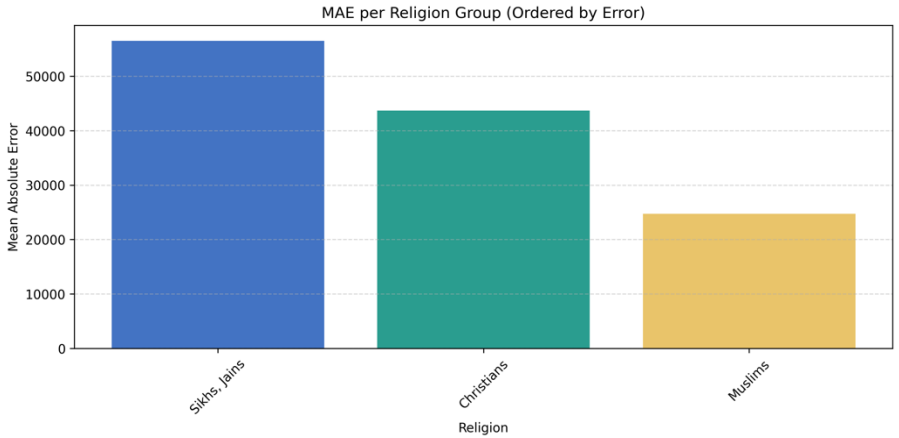
Sensitivity Analysis

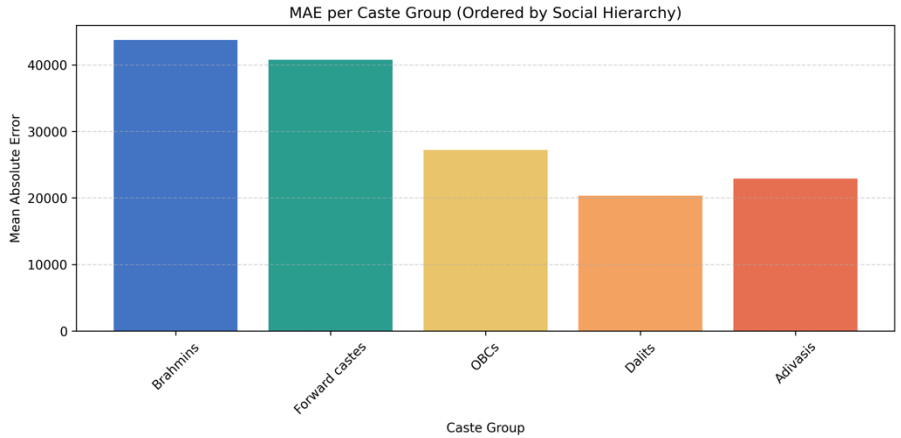


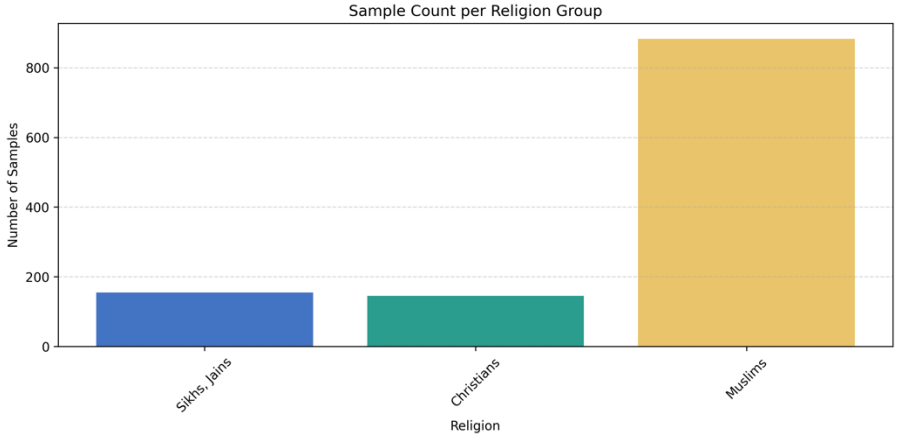
Bias and Fairness

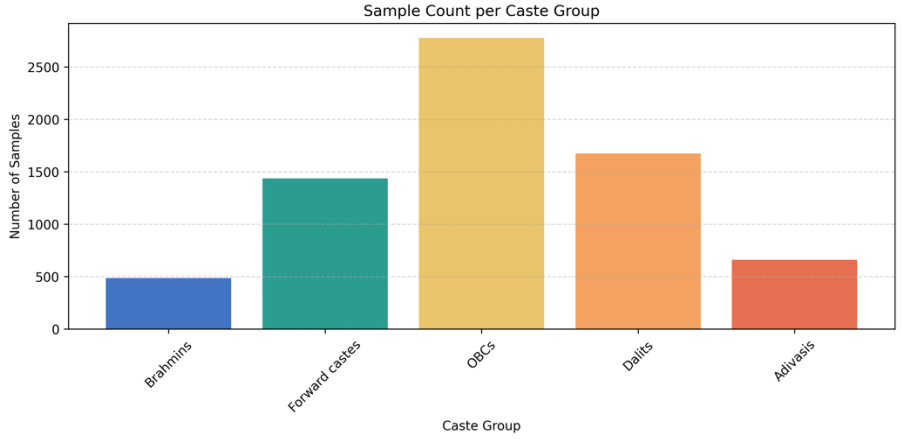


Bias and Fairness (cont.)









Conclusion

THOUGHTS

- **Neural network captures non-linear relationships** in the data that simple linear regression misses
- **Highest male education emerged as one of the most influential features**, indicating strong correlation between education and household income.
- **Fairness concerns**: Model performance varies significantly across caste an religion groups. (Bias or data imbalance)
- **Model struggles with extremes or outliers**: as observed in the residual analysis

IMPROVEMENTS

- **Focus on HH with <= 200’000 Rupee income** Try to improve model calibration and consider training a separate model for both segments.
- **Temporal information**
- **External data enrichment**
- **Feature engineering**

Summary

PERFORMANCE: R^2

US (aggregate)	Germany (aggregate)	US (individuals)	India (individuals)
~0.9	~0.7	~0.3	~0.3

CONCLUSION

- Aggregate income data are easier to predict than the income of individuals using socioeconomic factors (incl. education)
- Individuals exhibit more variance for similar backgrounds
- More factors are needed to know individual behaviour and environment important for predicition

→ Higher number of samples and features are needed, but unrealistic to receive such detailed and extensive datasets

Future Work

IMPROVEMENTS

- **Feature engineering**
- Compare prediction of more ML model

ADDITIONAL ANALYSIS

- Comparisons **intercontinental**:
  Use the same features and similar ML Models across countries to better assess the difference in feature importance across regions.
  e.g. predict income in USA for individual living in India

- **Agent-based modelling**
  to better understand interaction between individuals and explore emergent behaviour

Reference/ Literature/ Further Reading

SOURCES

1. ACS: 5-Year Estimates Data Profiles
2. Tiger: Shapefile

3. Regionaldatenbank Deutschland
4. Geobasis DE VG250 (Shapefile)

5. IPUMS US, custom dataset

6. IHDS
7. Shapefile (Github)

8. Github

LINK

1. https://data.census.gov/table?d=ACS+5-Year+Estimates+Data+Profiles
2. https://www2.census.gov/geo/tiger/

3. https://www.regionalstatistik.de/genesis/online
4. https://gdz.bkg.bund.de/index.php/default/digitale-geodaten/verwaltungsgebiete/verwaltungsgebiete-1-250-000-stand-01-01-vg250-01-01.html

5. https://www.icpsr.umich.edu/web/DSDR/studies/22626
6. https://github.com/AnujTiwari/India-State-and-Country-Shapefile-Updated-Jan-2020
7. IPUMS USA

8. https://github.com/simon-0006/Data-Science-GESS---Group-Project/tree/main

Questions?

Declaration of Originality

We hereby declare that this work is the result of our own original effort and has not been submitted, either in whole or in part, for academic credit at any other institution. All sources of information used in this work have been properly acknowledged, and all quotations are clearly indicated.

We affirm that each of us contributed to the work and are collectively responsible for its content. We understand the academic integrity policies of ETHZ and acknowledge the consequences of plagiarism or misrepresentation.

Contributions of Team Members

CONTENT	NAME
Country: US, Level: County	Pascal Gisiger
Country: Germany, Level: County	Erisa Ismaili
Country: India, Level: Household	Simon Ganter
Country: US, Level: Individual	Baban Deep Virk

Backup Slides- Germany

Possible Features

EMPLOYED RESIDENTS

MINIMUM INCOME SUPPORT BENEFITS

EMPLOYED PERSONS (AT PLACE OF WORK)

SCHOOLS (1995)

TARGET:

MEAN HOUSEHOLD INCOME

Possible Features

EMPLOYED RESIDENTS

- With an academic degree
- With recognized vocational degree
- Without vocational degree

(total and foreign)

MINIMUM INCOME SUPPORT BENEFITS

- Total standard benefit under SGB I (employable and non-employable)
- Assistance for living expenses
- Old age and disability
- Asylum Seekers Benefit

EMPLOYED PERSONS (AT PLACE OF WORK)

- Agriculture
- Industry
- Manufacturing
- Construction
- Trade & Transport
- Finance & Real Estate
- Public & Social Services

SCHOOLS (1995)

- Gymnasien
- Hauptschulen
- Realschulen
- Integrierte Gesamtschulen
- Sonderschulen / Förderschulen
- Schularten mit mehreren Bildungsgängen

TARGET:

MEAN HOUSEHOLD INCOME

Normalize by 15-65 years population within a Regional Unit

Possible Features

EMPLOYED RESIDENTS

- With an academic degree
- With recognized vocational degree
- Without vocational degree

(total and foreign)

MINIMUM INCOME SUPPORT BENEFITS

- Total standard benefit under SGB I (employable and non-employable)
- Assistance for living expenses
- Old age and disability
- Asylum Seekers Benefit

EMPLOYED PERSONS (AT PLACE OF WORK)

- Agriculture
- Industry
- Manufacturing
- Construction
- Trade & Transport
- Finance & Real Estate
- Public & Social Services

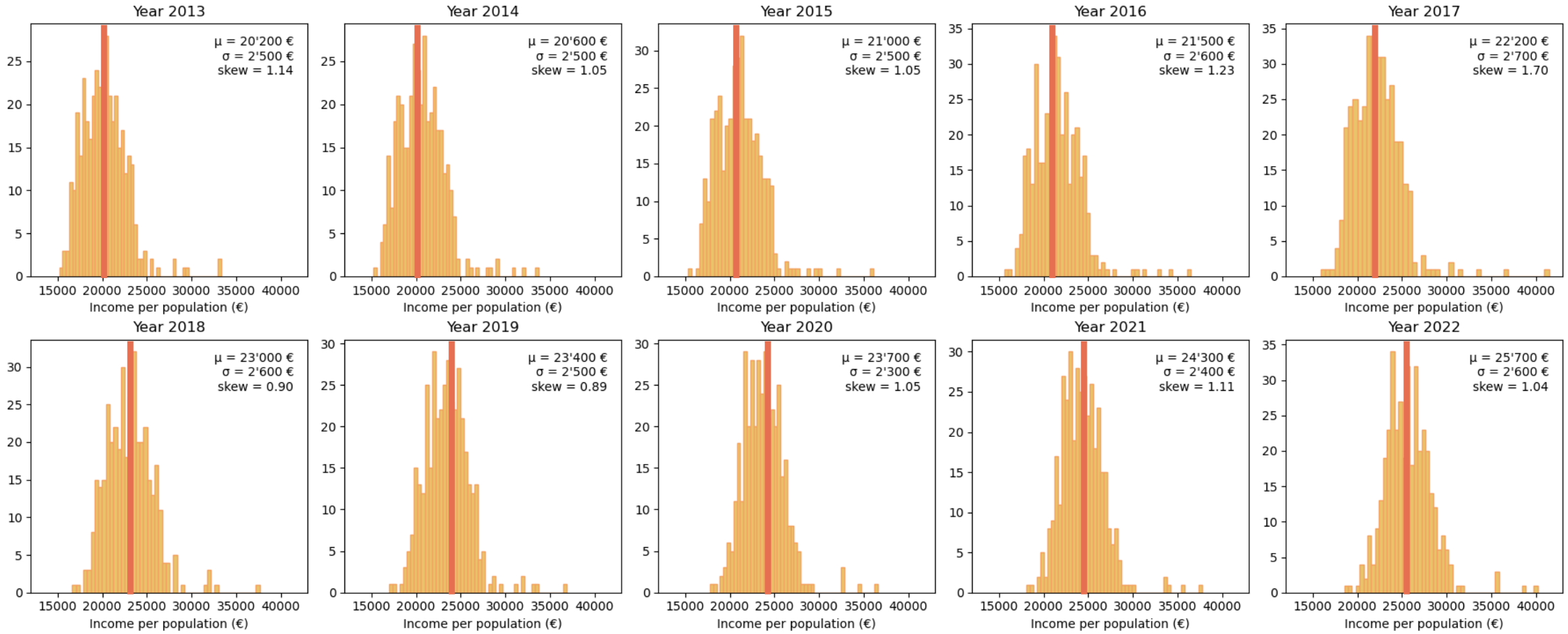
SCHOOLS (1995)

- Gymnasien
- Hauptschulen
- Realschulen
- Integrierte Gesamtschulen
- Sonderschulen / Förderschulen
- Schularten mit mehreren Bildungsgängen

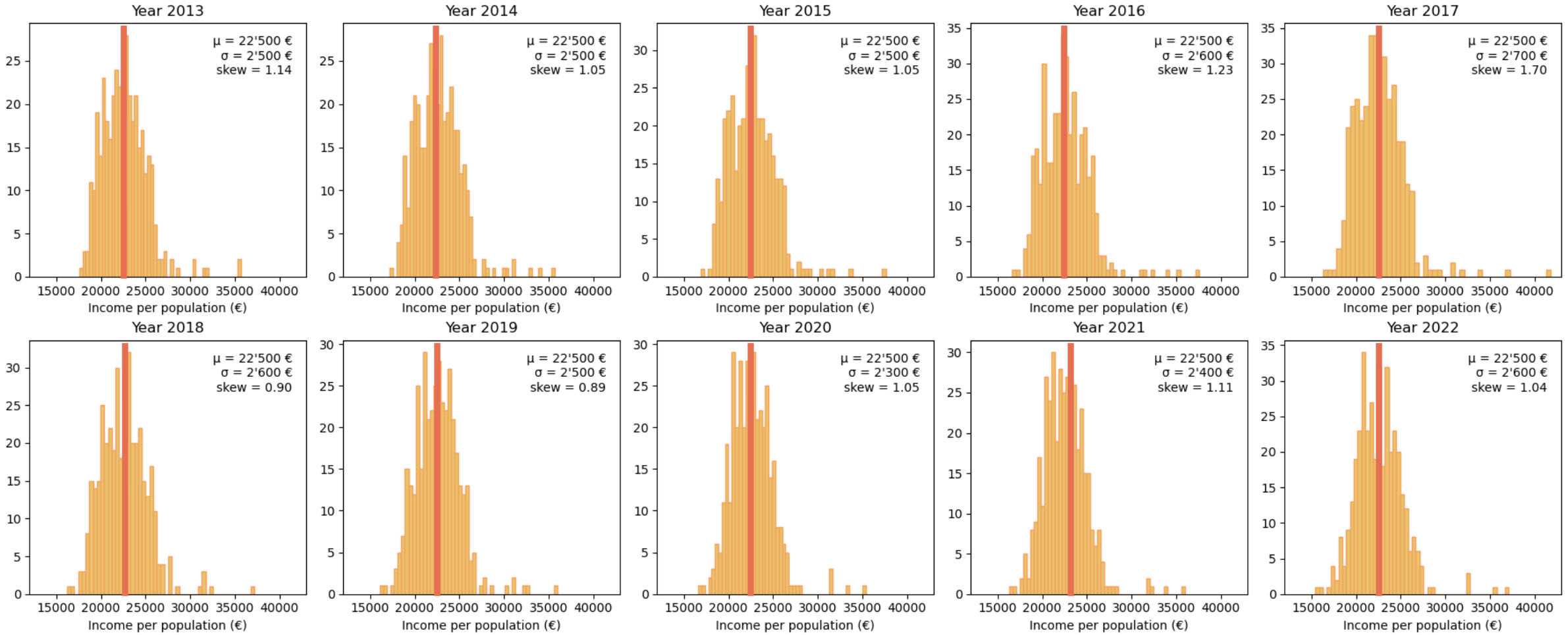
TARGET:

MEAN HOUSEHOLD INCOME

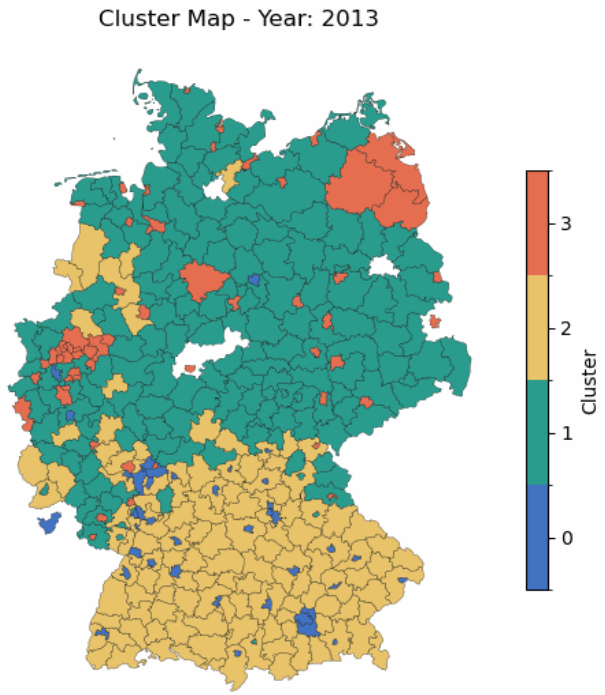
Income distribution



Income distribution (adjusted)

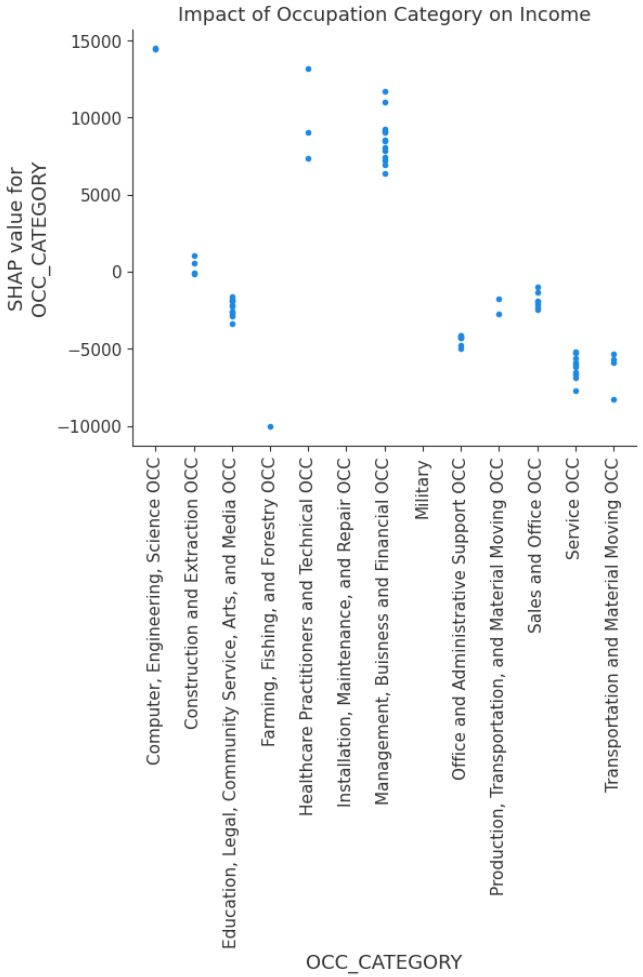


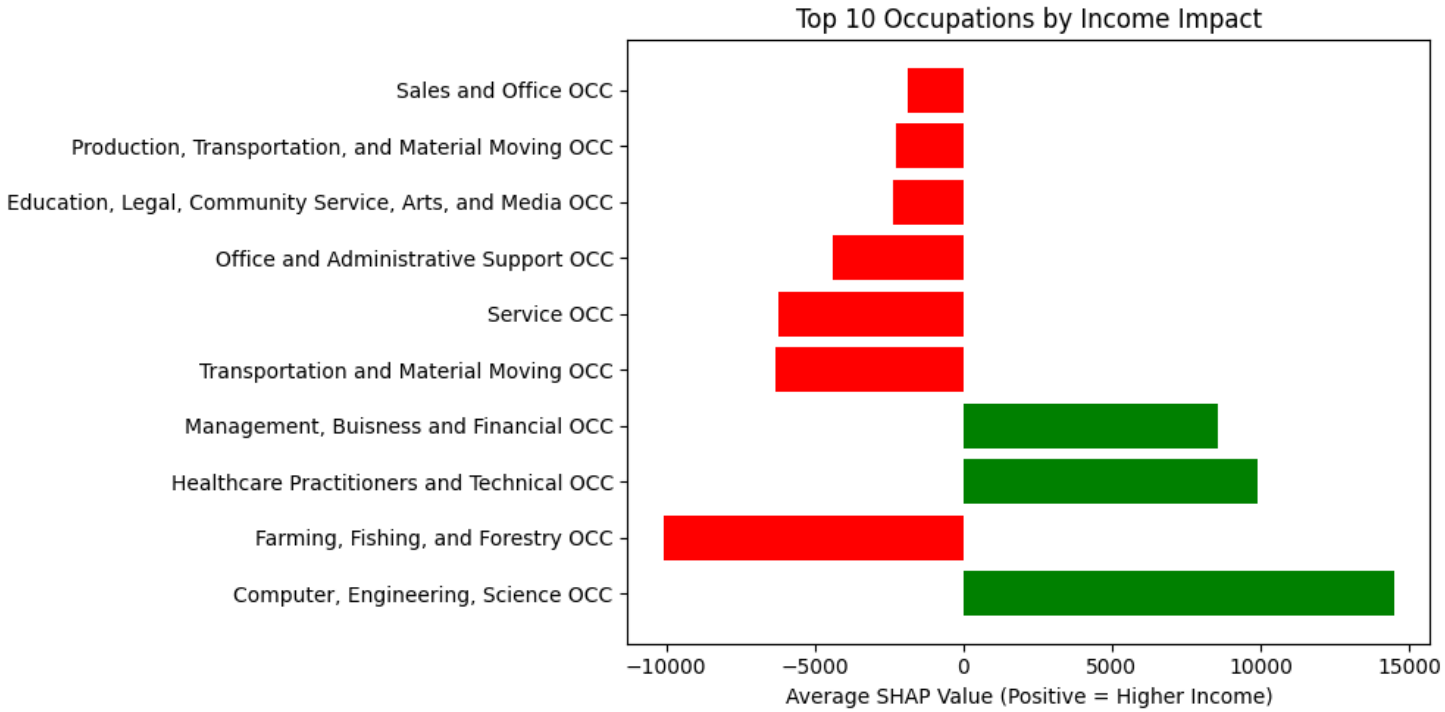
Clustering



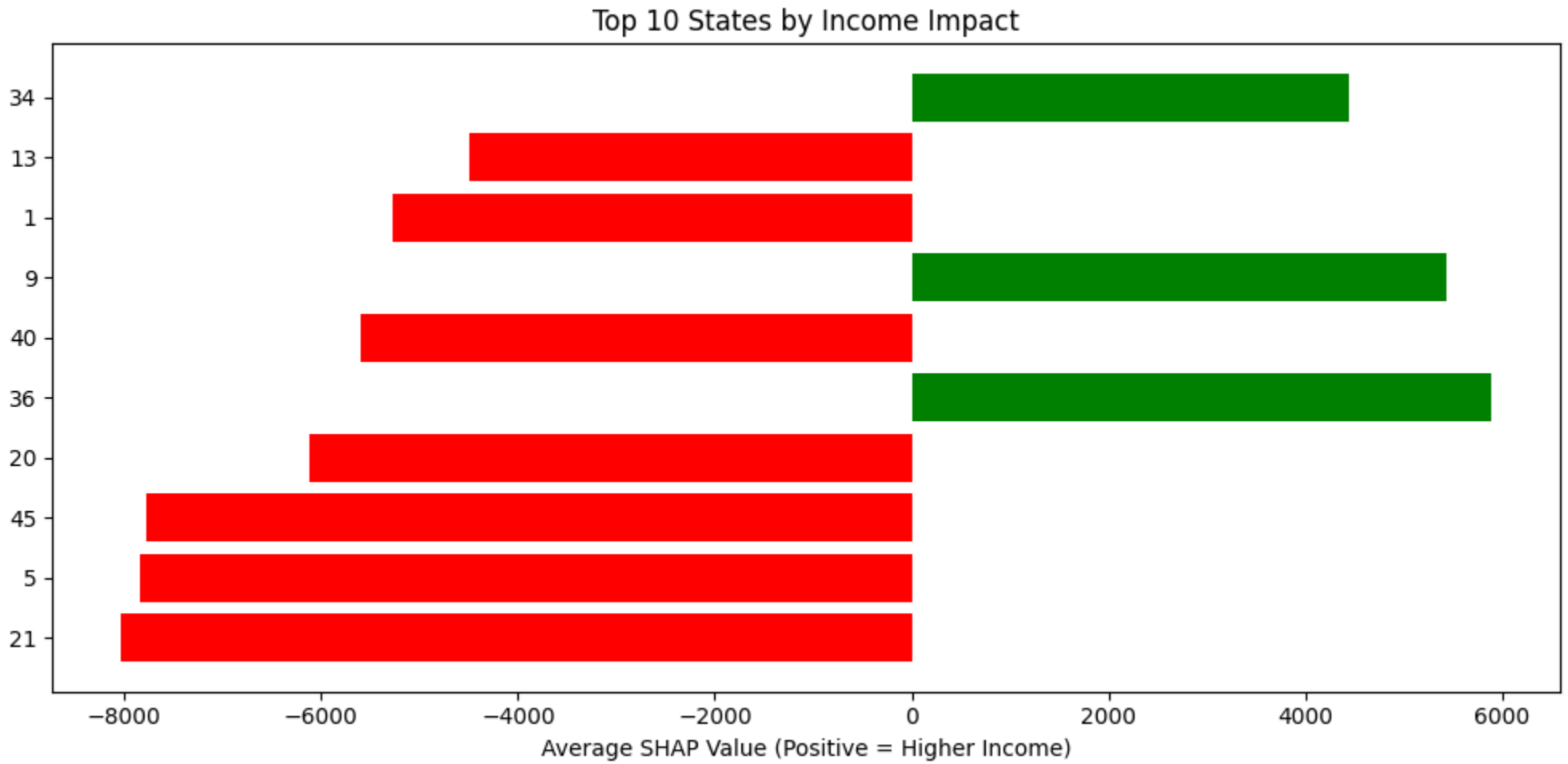
Backup Slides- US (individual)

INCOME BY OCC

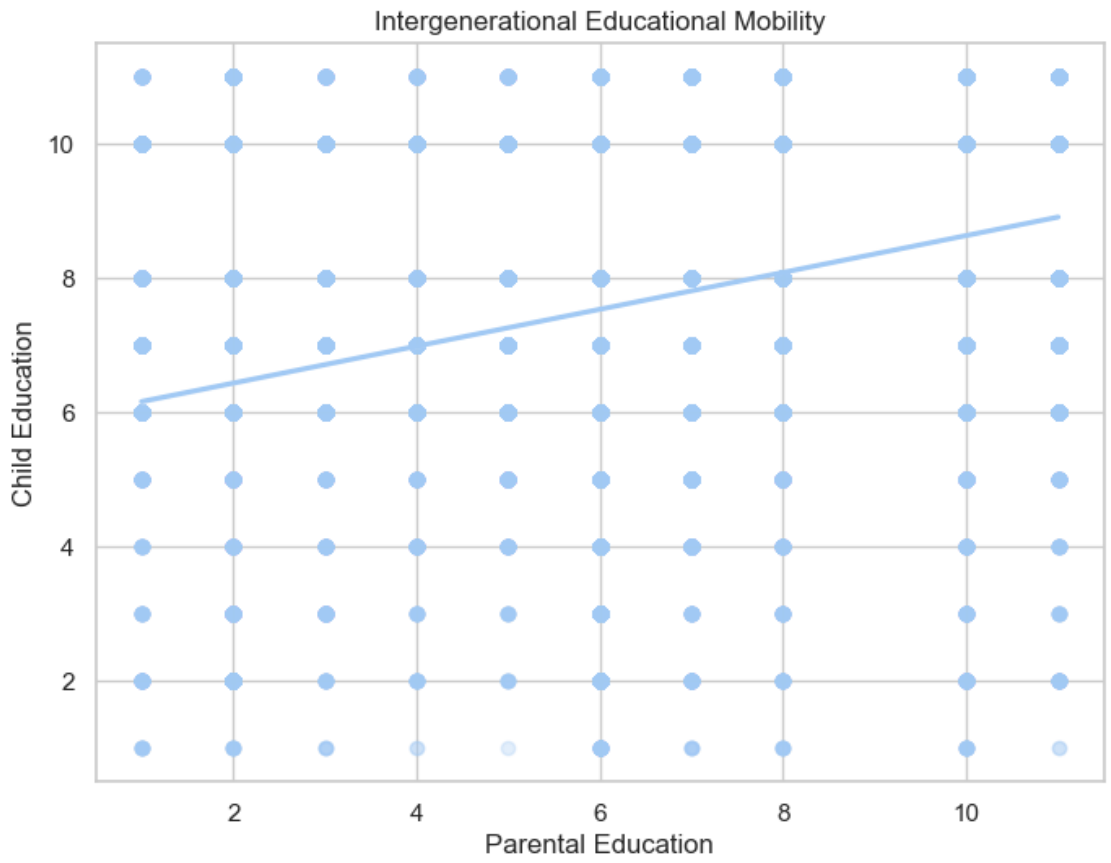


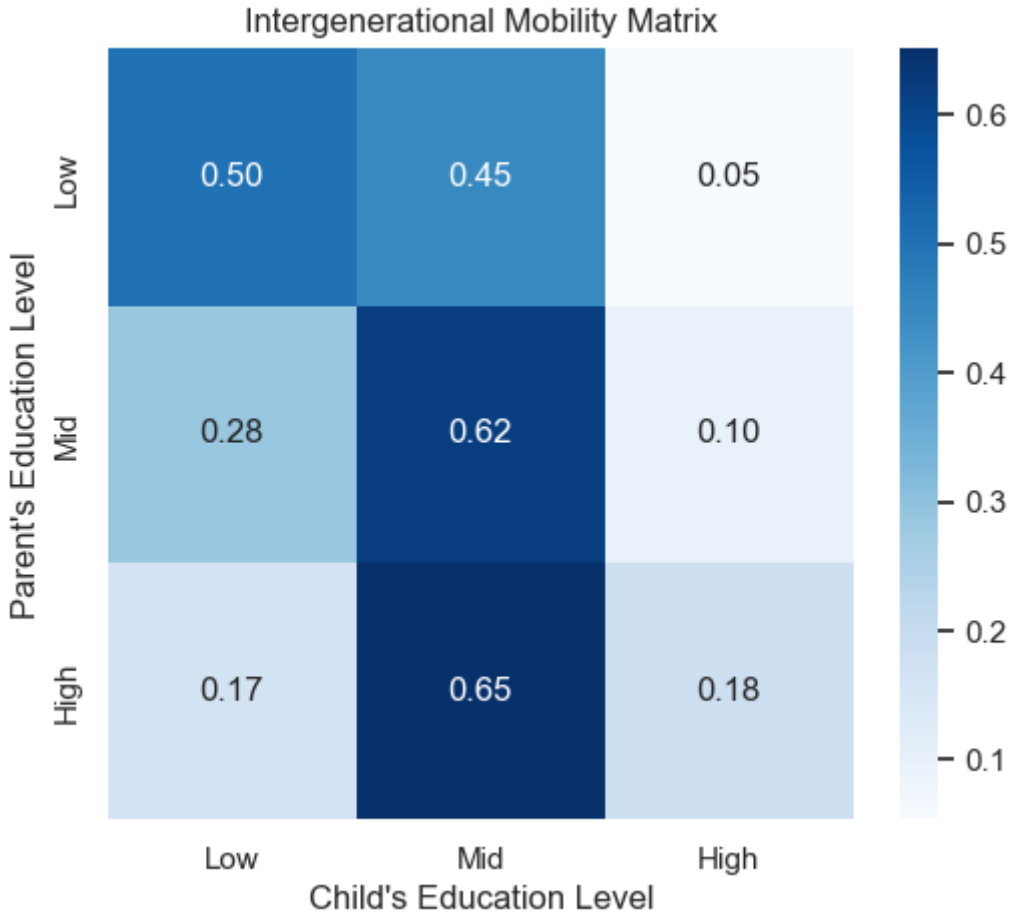


INCOME BY STATE



CHILD/PARENT EDUC





CHILD/PARENT EDUC