Strategic Thinking

Computation of quality life index for world countries

Team 3:
Simone Finelli, Thomas Kelly, Faraj Alheajneh and Akash Jyothis Parappurathu

Lecturer:
James Garza

Business Understanding

The objective of this project is to predict the value of a potential income based on cost of commodities as variables to predict quality of life in a given city.

Hypothesis

- The cost of living in a country is a significant factor in determining the quality of life.
- Nonfinancial factors such as healthcare, education, safety, and environmental quality contribute significantly to the overall quality of life in a country.
- Countries with higher price level indices tend to have a lower quality of life compared to countries with lower price level indices.

General goal

- To predict the cost of living and the “quality of life index” in a given year in a given city based on the potential income

Success criteria/indicators

- Be able to predict with a 90% accuracy the country's quality of life index.

Business Understanding

The requirements for this project are:

- Data preparation that gives support to investigate the data and prepares it for the machine learning model;
- Data on cost of living. We will need access to reliable and up-to-date data on the cost of living in different countries. This can include information on housing, transportation, food, utilities, and other expenses. Sources like official government statistics can provide relevant data.
- Data on nonfinancial factors. We would need data on various non-financial factors that contribute to the quality of life in a country. This can include factors such as healthcare quality, education system, safety, environmental.

Data Understanding PPP

Dataset

- “How many currency units a given quantity of goods and services costs in different countries”, available on Eurostat. Yearly data from 1995 to 2021.

Shape

- 898, 14.

Attributes

- **Geo**, **Time**, **A0101**, **A0102**, **A0103**, **A0104**, **A0105**, **A0106**, **A0107**, **A0108**, **A0109**, **A0110**, **A0111**, **A0112**.

	Unnamed: 0.1	Unnamed: 0	DATAFLOW	LAST UPDATE	freq	na_item	ppp_cat	geo	TIME_PERIOD	OBS_VALUE	OBS_FLAG
0	0	0	ESTAT:PRC_PPP_IND(1.0)	24/06/22 23:00:00	A	EXP_EUR	A0101	AL	2005	1945.0	NaN
1	1	1	ESTAT:PRC_PPP_IND(1.0)	24/06/22 23:00:00	A	EXP_EUR	A0101	AL	2006	2094.0	NaN
2	2	2	ESTAT:PRC_PPP_IND(1.0)	24/06/22 23:00:00	A	EXP_EUR	A0101	AL	2007	2318.0	NaN
3	3	3	ESTAT:PRC_PPP_IND(1.0)	24/06/22 23:00:00	A	EXP_EUR	A0101	AL	2008	2606.0	NaN
4	4	4	ESTAT:PRC_PPP_IND(1.0)	24/06/22 23:00:00	A	EXP_EUR	A0101	AL	2009	2599.0	NaN

PPP dataset head

geo	TIME_PERIOD	A0101	A0102	A0103	A0104	A0105	A0106	A0107	A0108	A0109
AL	2005	10798.578962	1067.136188	1258.383208	4233.016204	1842.115858	1922.970629	1128.810788	584.973458	715.892221
AL	2006	11524.835975	1136.265175	1340.608250	4637.742171	1963.007629	2001.091329	1254.309925	603.739458	767.773242
AL	2007	12813.186767	1256.999012	1627.168083	5059.319254	2635.215342	2263.231475	1667.941146	720.898625	1021.797979
AL	2008	14324.853017	1411.864467	1810.864708	5400.331463	2948.154275	2552.566662	1893.218654	879.877292	1140.930946
AL	2009	15304.002433	1445.891029	1780.601208	5488.864742	2862.185421	2880.220850	1836.851571	906.190233	1136.259129

PPP dataset pivot

Data Understanding PPP

Features decoding

- A0101 -> Food and non-alcoholic beverages
- A0102 -> Alcoholic beverages, tobacco and narcotics
- A0103 -> Clothing and footwear
- A0104 -> Housing, water, electricity, gas and other fuels
- A0105 -> Household furnishings, equipment and maintenance
- A0106 -> Health
- A0112 -> Miscellaneous goods and services
- A0107 -> Transport
- A0108 -> Communication
- A0109 -> Recreation and culture
- A0110 -> Education
- A0111 -> Restaurants and hotels

Data Understanding PLI

Dataset

- “A measure of the differences in the general price levels of different countries”, available on Organisation for Economic Co-operation and Development (OECD). Yearly data from 1997 to 2021.

Shape

- 1175, 8.

Attributes

- **Location**, Indicator, Subject, Measure, Frequency, **Time**, **Value**, Flag Codes.

	LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value	Flag Codes
0	AUS	PLI	TOT	OECDIDX	A	1997	97	NaN
1	AUS	PLI	TOT	OECDIDX	A	1998	84	NaN
2	AUS	PLI	TOT	OECDIDX	A	1999	86	NaN
3	AUS	PLI	TOT	OECDIDX	A	2000	82	NaN
4	AUS	PLI	TOT	OECDIDX	A	2001	77	NaN

PLI dataset head

Data Preparation

What we did:

- Named countries in a coherent way
- Merged datasets
- No ‘na’ or ‘nan’ values
- Mapped an index to each country
- Scaled data
- Removed outliers

Country naming

```
In [26]: np.unique(pli["LOCATION"])

Out[26]: array(['AUS', 'AUT', 'BEL', 'BRA', 'CAN', 'CHE', 'CHL', 'CHN', 'COL',
                'CRI', 'CZE', 'DEU', 'DNK', 'EA18', 'ESP', 'EST', 'EU27', 'FIN',
                'FRA', 'GBR', 'GRC', 'HUN', 'IDN', 'IND', 'IRL', 'ISL', 'ISR',
                'ITA', 'JPN', 'KOR', 'LTU', 'LUX', 'LVA', 'MEX', 'NLD', 'NOR',
                'NZL', 'OECD', 'POL', 'PRT', 'RUS', 'SVK', 'SVN', 'SWE', 'TUR',
                'USA', 'ZAF'], dtype=object)
```

Countries available in PLI dataset

```
In [27]: np.unique(ppp["geo"])

Out[27]: array(['AL', 'AT', 'BA', 'BE', 'BG', 'CH', 'CY', 'CZ', 'DE', 'DK', 'EE',
                'EL', 'ES', 'FI', 'FR', 'HR', 'HU', 'IE', 'IS', 'IT', 'LT', 'LU',
                'LV', 'ME', 'MK', 'MT', 'NL', 'NO', 'PL', 'PT', 'RO', 'RS', 'SE',
                'SI', 'SK', 'TR', 'UK', 'XK'], dtype=object)
```

Countries available in PPP dataset

Dataset merging

```
In [35]:  # column renaming for PPP

          ppp.rename(columns = {"geo":"LOCATION", "TIME_PERIOD":"TIME"}, inplace = True)
```

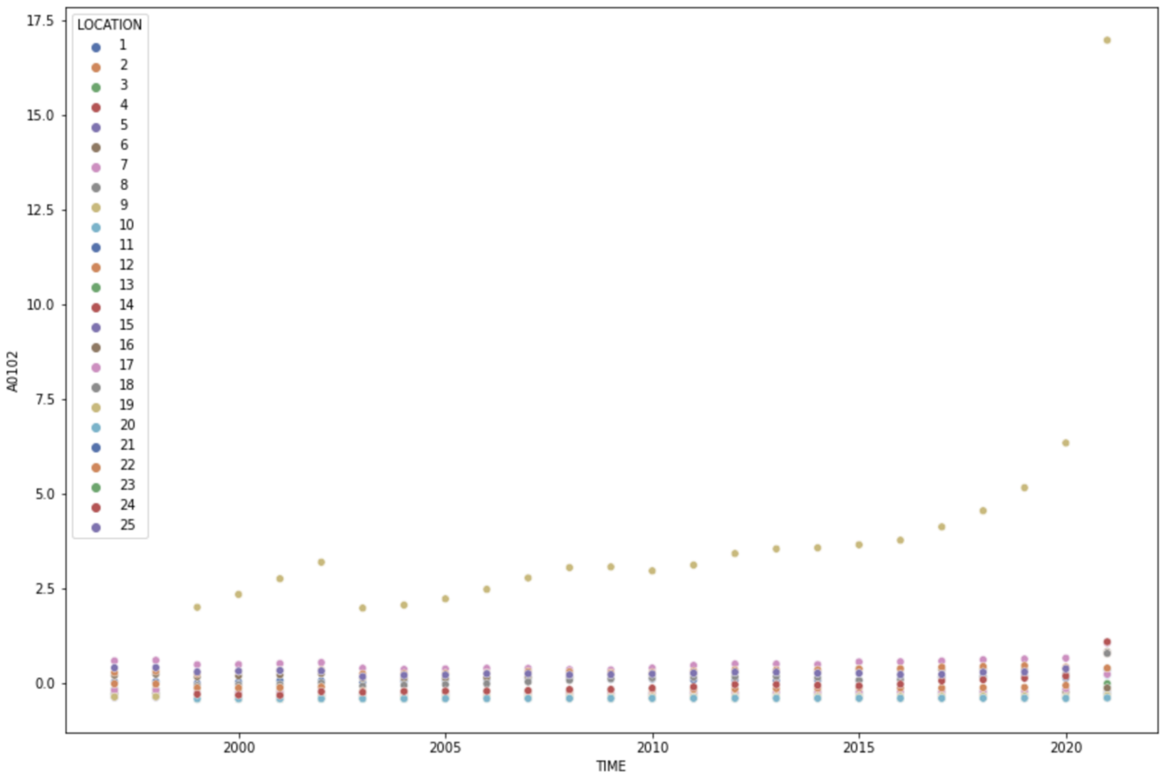
Countries renaming in PLI

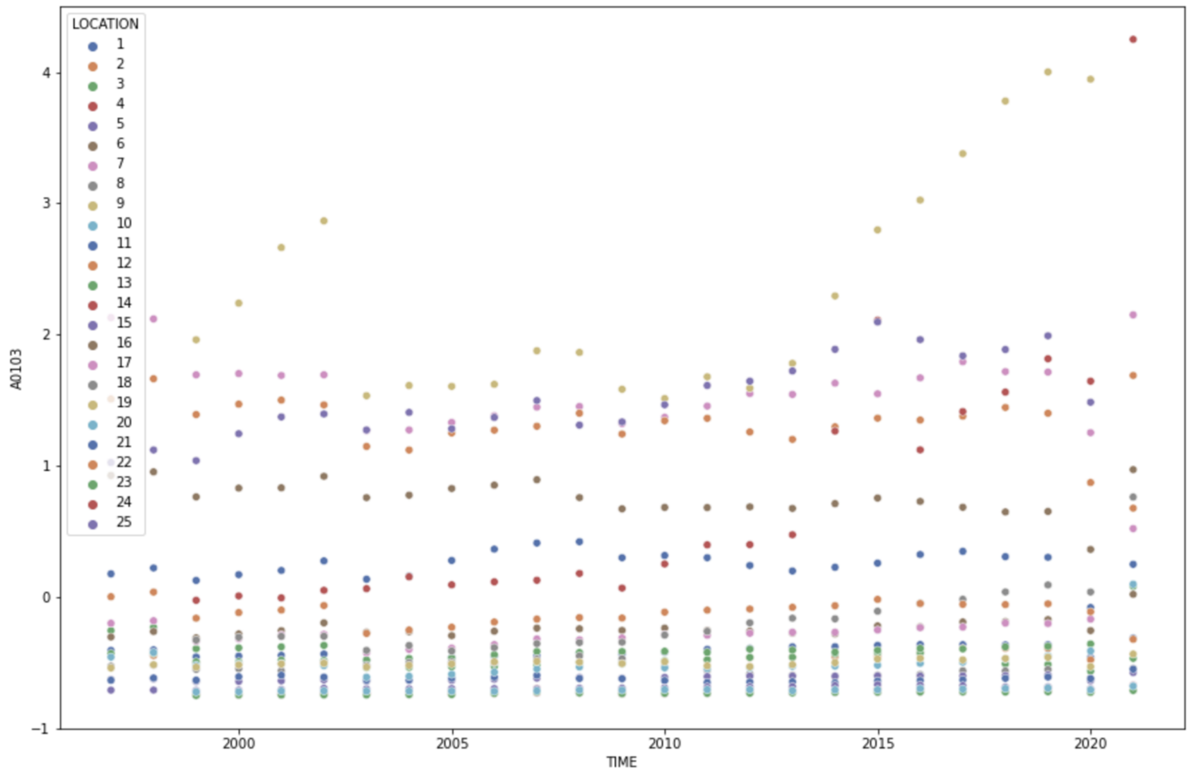
```
In [36]:  # PPP and PLI outer merge on LOCATION and TIME columns

          df = pd.merge(ppp, pli, how="outer", on=["LOCATION", "TIME"])
```

Datasets outer merge

Final number of samples: **610**

Outliers removal





Country **Hungary** was dropped due to evident outliers

Country indexing and data scaling

	LOCATION	TIME	A0101	A0102	A0103	A0104	A0105	A0106	A0107	A0108	A0109	A0110	A0111	A0112
0	1	1997	-0.394902	-0.334167	-0.406500	-0.471679	-0.390710	-0.460854	-0.462510	-0.445156	-0.453544	-0.453619	-0.368401	-0.517210
1	1	1998	-0.394116	-0.328095	-0.402795	-0.466103	-0.389824	-0.446995	-0.457922	-0.435548	-0.441328	-0.445955	-0.343586	-0.507590
2	1	1999	-0.412964	-0.338755	-0.456795	-0.482730	-0.419292	-0.468856	-0.478547	-0.443756	-0.467238	-0.480575	-0.376576	-0.531943
3	1	2000	-0.408245	-0.335875	-0.450570	-0.474204	-0.405953	-0.461238	-0.466331	-0.392574	-0.454384	-0.473952	-0.358806	-0.519730
4	1	2001	-0.403680	-0.334179	-0.444267	-0.467774	-0.409297	-0.454624	-0.466914	-0.392603	-0.448108	-0.468429	-0.361660	-0.518293
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
605	25	2016	0.366294	0.228845	1.961140	1.047494	0.831352	0.941282	1.180932	0.417685	2.104035	0.933012	1.362254	1.196904
606	25	2017	0.390496	0.233735	1.837536	1.032374	0.910785	0.941184	1.185175	0.436171	2.154815	0.955515	1.462471	1.195337
607	25	2018	0.427729	0.289610	1.885649	1.081231	1.042757	0.962788	1.314897	0.702938	2.134240	0.983847	1.515223	1.349229
608	25	2019	0.468716	0.297004	1.990590	1.121580	1.171777	1.082541	1.382596	0.669644	2.268071	1.106124	1.543097	1.387723
609	25	2020	0.498634	0.381739	1.484233	1.104352	1.218920	1.261537	0.655448	0.512336	1.967289	1.108724	0.315890	1.169533

Table showing country indexes and scaled data

Modeling

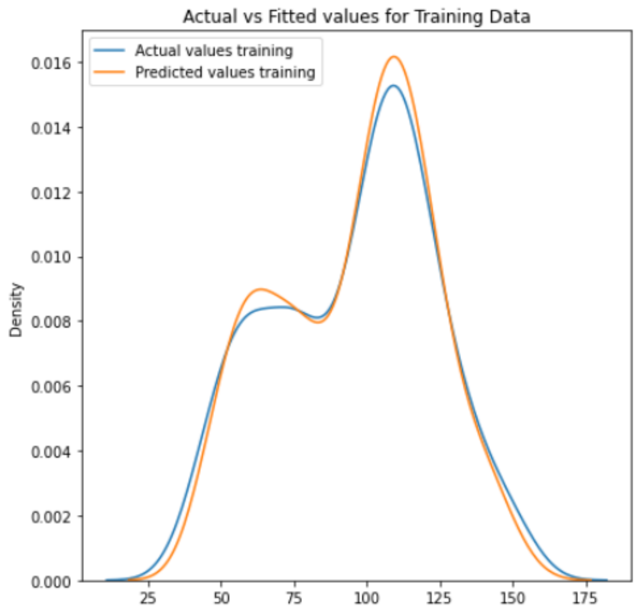
Steps

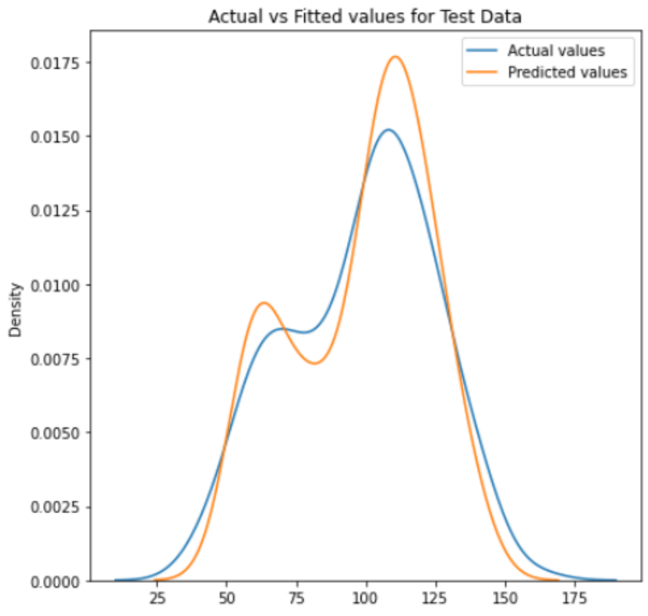
- Splitting the data into training and test sets
- Tested all models without scaling and optimizations
- Tested all models with dataset scaling
- Tested all models with optimizations using GridSearchCV

Methodology

- Continuous evaluation with MSE and R2 metrics
- Models built using Random Forest, XGBoost and Neural Networks

Random Forest





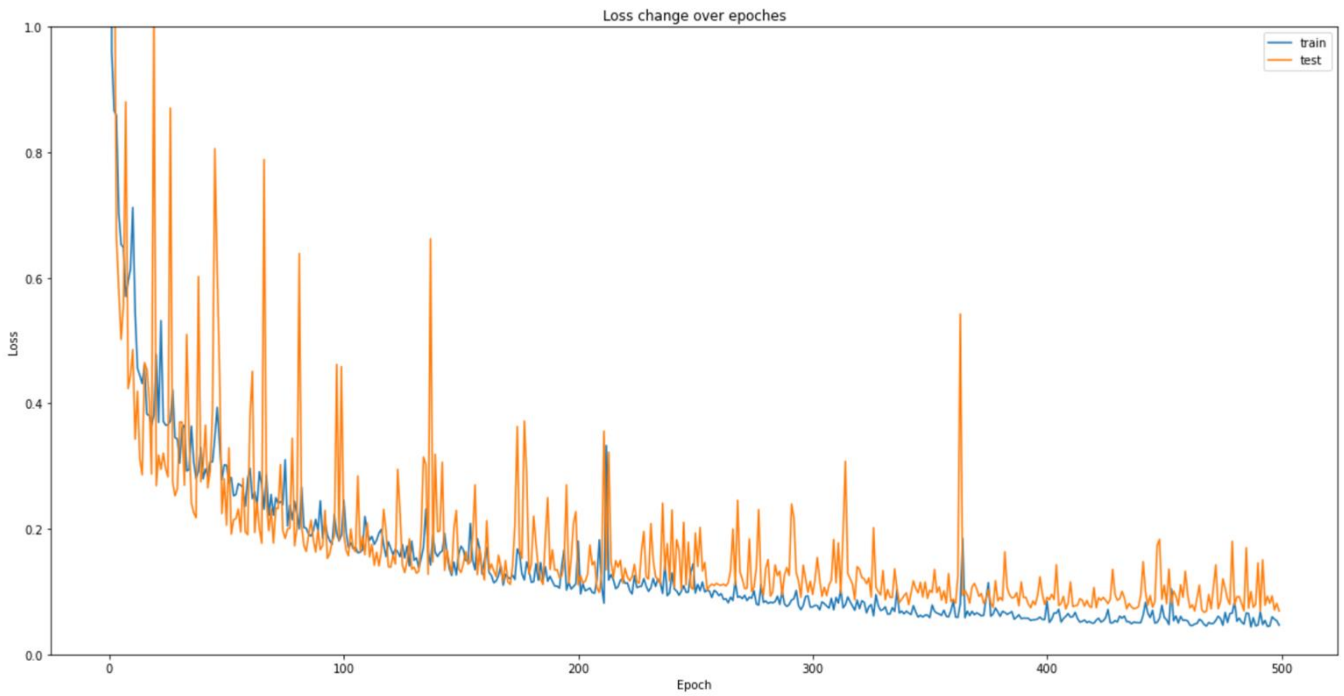
Outcome

- MSE improvement with scaling
- Clear overfitting
- GridSearchCV no improvement

```
In [110]: model.best_params_

Out[110]: {'max_depth': 10,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 200}
```

	RFR - No optimization		**RFR - Dataset scaling**		**RFR – GridSearchCV**	
	Training	*Test*	*Training*	*Test*	*Training*	*Test*
R2 Score	0.986	0.906	0.984	0.905	0.986	0.907
MSE Score	10.665	64.002	0.016	0.086	0.015	0.084

Neural Network



Outcome

- Noise in loss
- Divergence from 250th epoch
- Good results

Init mode: Normal

Optimizer: Adagrad

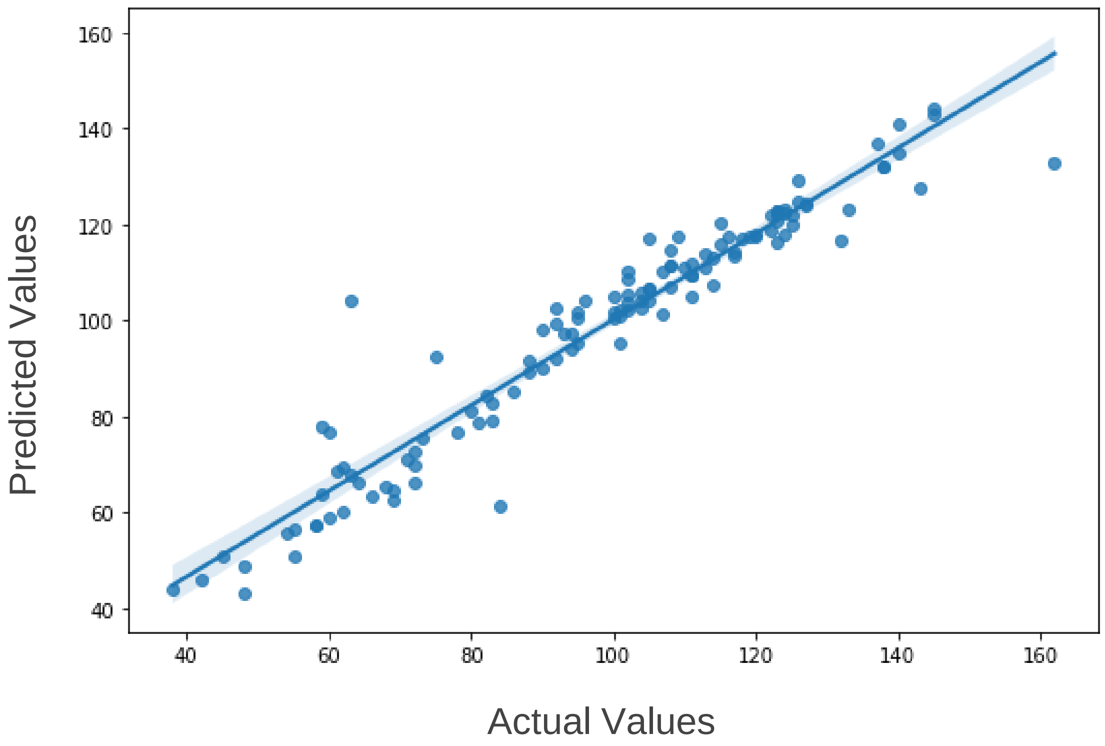
Learn rate: 0.3

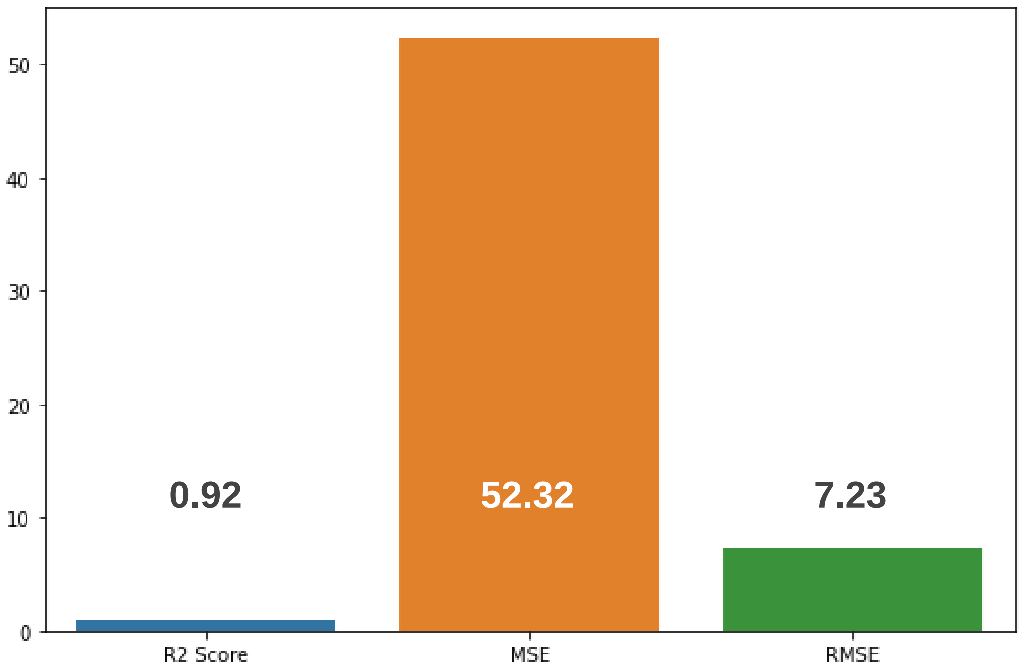
neurons: 50

epochs: 500

	NN - No optimization		NN – GridSearchCV	
	Training	*Test*	*Training*	*Test*
R2 Score	0.755	0.743	0.958	0.923
MSE Score	0.255	0.233	0.0043	0.069

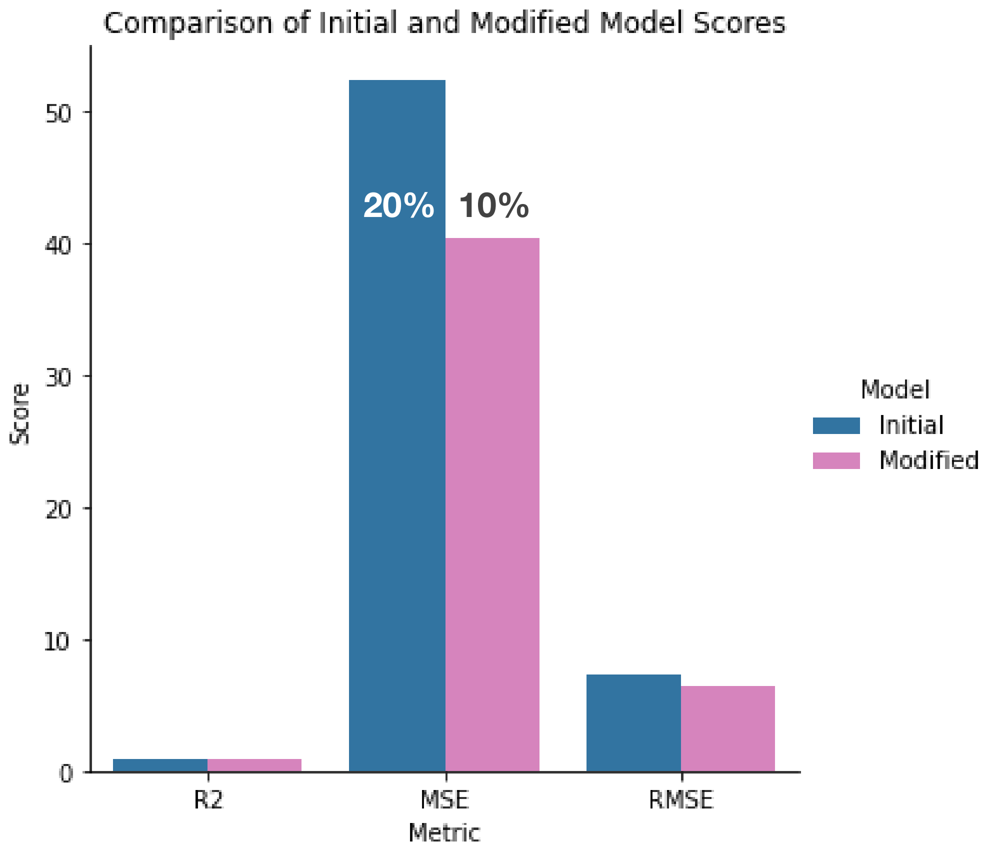
XGBoost – Untuned





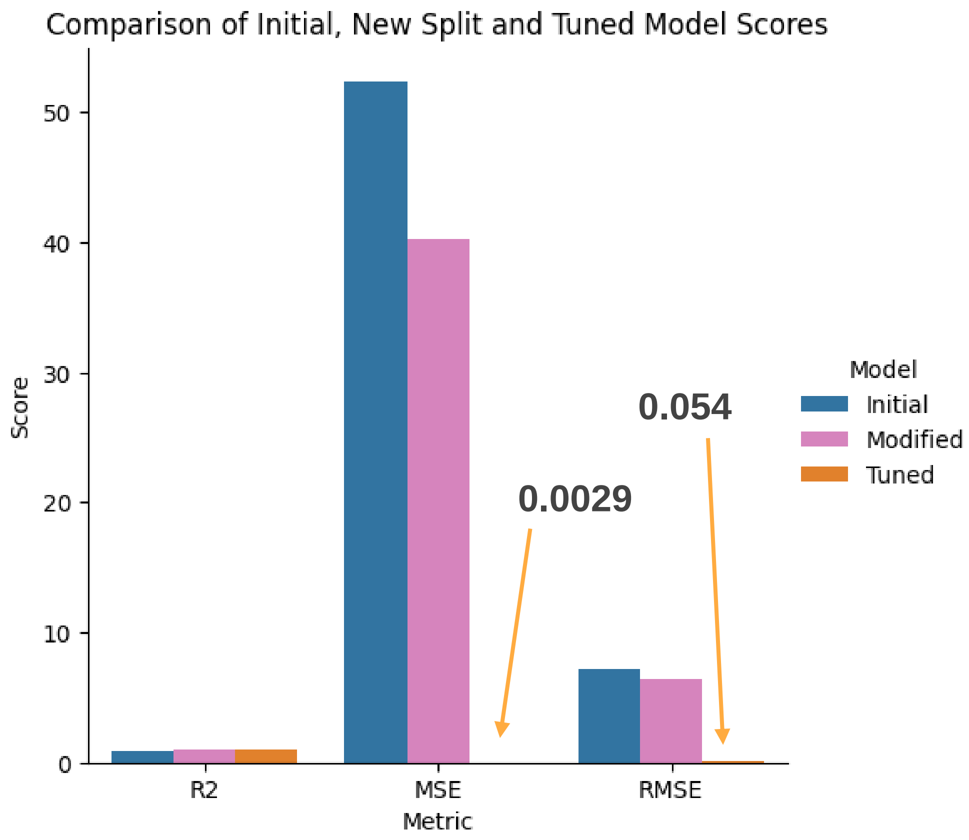
XGBoost – Tuning

Changing Splits



GridSearchCV

- ‘colsample_bytree’: 1
- ‘learning_rate’: 0.2
- ‘max_depth’: 3
- ‘min_child_weight’: 3
- ‘subsample’: 0.8

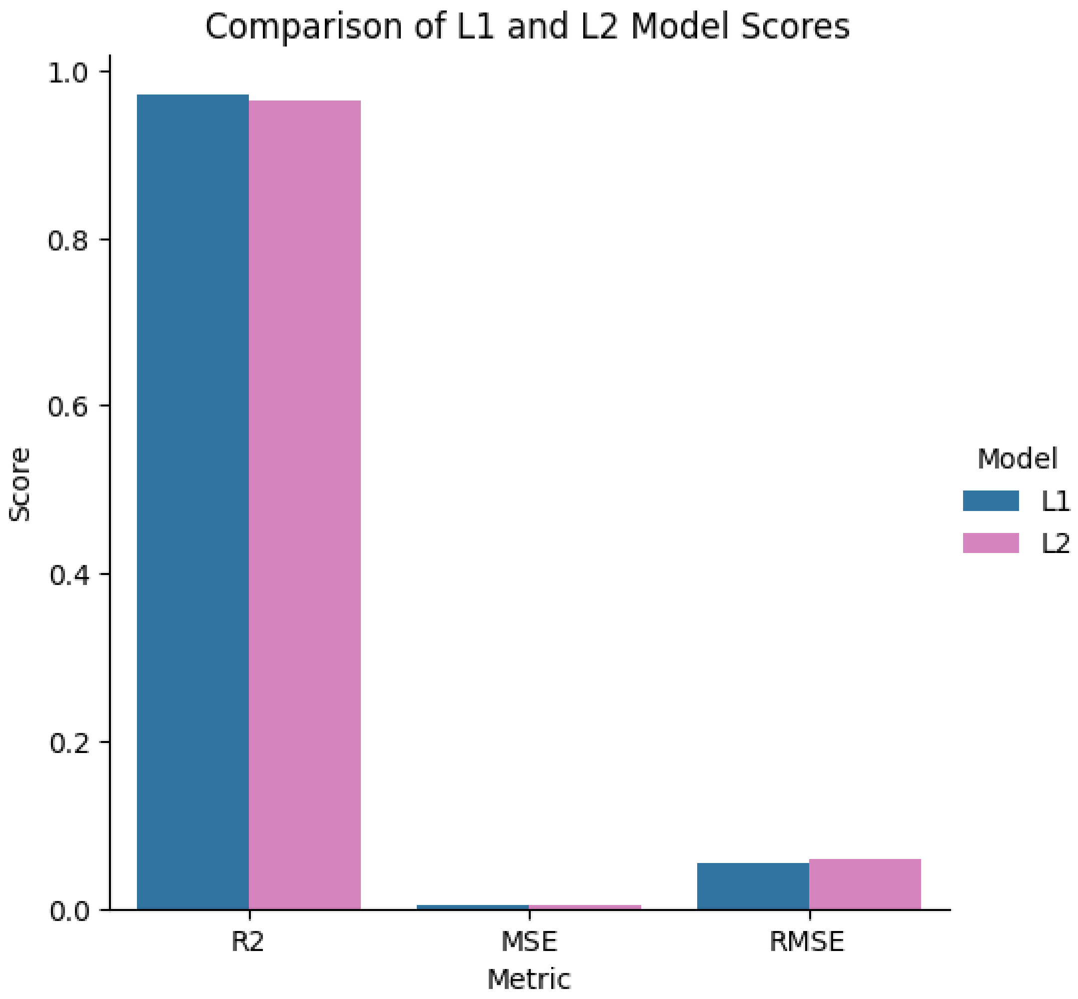


XGBoost – Tuning

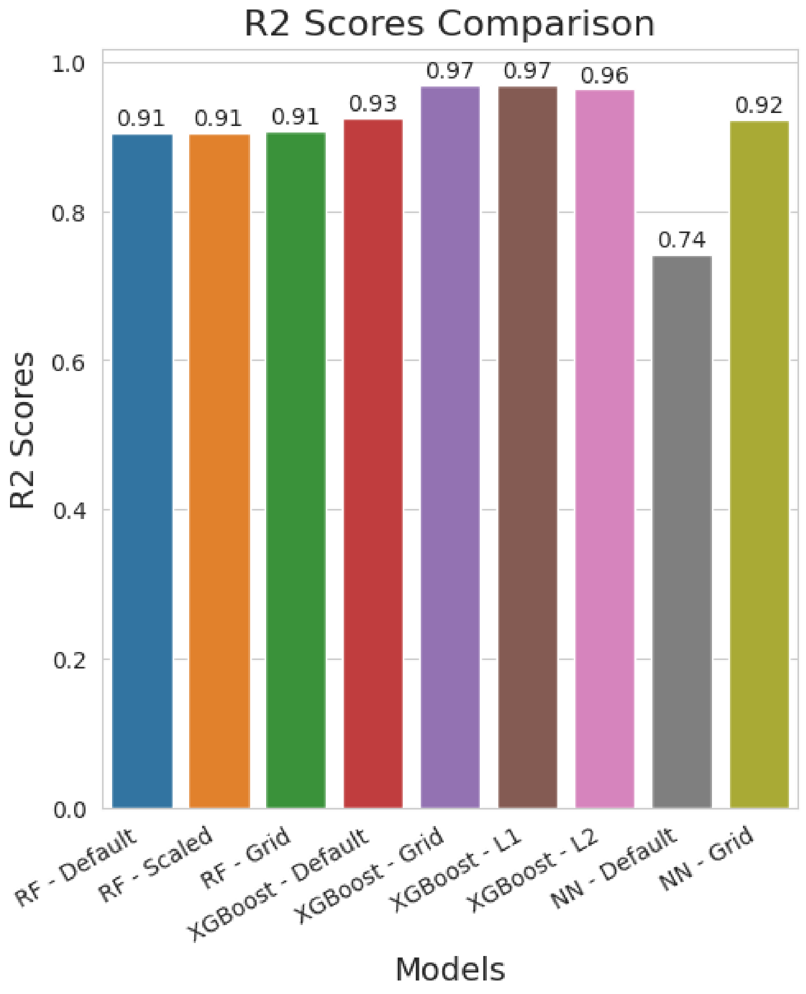
L1 & L2 Regularisation

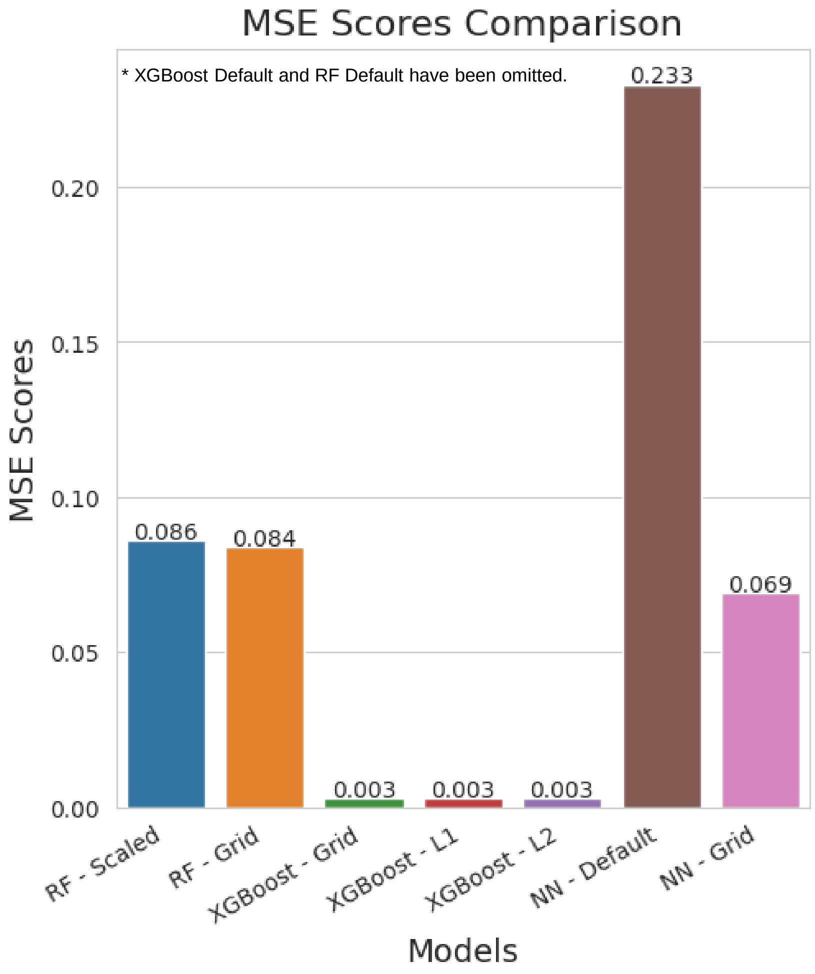
```
# L1 Regularisation
model = xgb.XGBRegressor(objective='reg:squarederror', alpha=0.1, **best_params)
```

```
# L2 Regularisation
model = xgb.XGBRegressor(objective='reg:squarederror', reg_lambda=0.1, **best_params)
```



Modeling comparison





1. **XGBoost Tuned (L1)**
2. **Neural Network (Tuned)**
3. **Random Forest (Tuned)**

Conclusion and recommendations

Success criteria was met.

Reflections and next steps for this project would be:

- Find new way to give numeral values to non-monetary data such as political and policies in the countries to have a better factors for the MLM to compute a more accurate prediction.
- Include multi dimensions of data that can have a reflection on quality of life, pension data for example.
- Avoid bias and generalization.
- Be city specific rather than country for more precise results.
- Use of cloud computing to shape a high accuracy tool and detect data drift.