UNIVERSITÄT BERN

Mathematical Institute

CAS Applied Data Science

Income Modeling Based On Socioeconomic Factors

Applying Machine Learning Algorithms to Registry and Tax Data from the Canton of Lucerne

David von Holzen
Ziegeleiweg 1, 6048 Horw
david.vonholzen@bluewin.ch

June 13, 2025

Abstract

Machine learning algorithms offer powerful potential for data analysis in statistical offices. One field in which machine learning algorithms are applied, and a statistical office could potentially benefit from, is prediction modeling. The register statistics of LUSTAT, a dataset which combines tax and register data on an individual level, provides a promising opportunity to try out different machine learning algorithms for prediction modeling. It contains tax data from 2010 to 2022 and register data from 2010 to 2024 from the Canton of Lucerne. Using this dataset, the taxable income for the years 2023 and 2024 is estimated from socioeconomic factors in the register data (e.g., age, gender & nationality). A linear regression model, a random forest model, a gradient boosting model and a recurrent neural network (RNN) are built up for income prediction. The models are trained and tested on the data from 2010 to 2022. Due to data privacy restrictions, the observations are spatially aggregated into raster cells with a minimum of 10 observations. Consequently, the feature and target variables are also aggregated on the level of these raster cells. The comparison of the performance of the four machine learning algorithms shows, that the RNN-model provides the best income prediction. The gradient boosting model provides the second best prediction. The linear and the random forest model show the weakest performance. Subsequently, the gradient boosting model is improved by integrating sequential information. Despite the improvements, the model does not reach the performance level of the RNN-model. However, even the RNN-model is not able to produce plausible income predictions. The aggregation of variables at the raster cell level likely results in a loss of information and variability, which complicates the modeling process. As a consequence, a final evaluation of the tested machine learning algorithms for use in a statistical office is not possible. Therefore, as a next step, the models should be trained and tested on individual-level data.

Contents

1 Introduction

Machine learning algorithms offer powerful potential for data analysis in statistical offices. One field in which machine learning algorithms are applied, and a statistical office could potentially benefit from, is prediction modeling. However, prediction does not necessarily refer to the future but also the estimation of values which are in the past but their collection is not carried out yet. Hence, a pivotal challenge for a statistical office is the time lag concerning the availability of the necessary data. For example, tax data entails a long process from filling out the tax declaration by the citizens until having them prepared as a dataset at the statistical office. On the contrary, data about the residents and the buildings, so-called register data, offer a high degree of timeliness. This opens up the opportunity to use more current data sources to model data that is not available yet. The register statistics of LUSTAT, a dataset which combines tax and register data on an individual level, currently contains tax data from 2010 to 2022 and register data from 2010 to 2024. This analysis tries to contribute to the challenge of time-delayed data availability by estimating the taxable income with the help of certain socioeconomic factors (e.g. age, gender & nationality). More specifically, a linear regression model, a random forest model, a gradient boosting model and a recurrent neural network will be built up based on the data from 2010 to 2022 to predict the income in 2023 and 2024. On a more general level, the analysis offers the possibility to test and compare the applicability of the different machine learning algorithms for a potential use at the statistical office.

The following paper is organized as follows: In the first part, the data preparation process is described. The second section takes a look at the distribution of the feature and target variables. In the third part, different machine learning algorithms are applied to train and test models for income prediction. In the fourth part, the results of the machine learning analysis are discussed. In the final section, conclusions are drawn and the findings are used to take a look in the future, i.e. possible applications of the findings for the statistical office are outlined.

2 Data

2.1 Data Preparation

The dataset is based on the registry statistics of LUSTAT, the statistical office of the Canton of Lucerne, and contains information from the register of residents and buildings and from the tax statistics of the Canton of Lucerne. Currently, the register data reaches from 2010 to 2024 and the tax data from 2010 to 2022. The dataset is stored in the in-house data warehouse and can be accessed by SAS Enterprise Guide or the Posit Workbench and therefore does not have

to be collected for this project. Since it is also a well prepared and stored dataset, the data already possesses a high quality. Due to data privacy restrictions the data cannot be used on the individual level. The dataset is in a first step filtered by the necessary variables as the age of the individuals, only observations of working age are used (18-64 years), and the household type, only observations in private households are used. Furthermore, due to tax data privacy restrictions, only observations with a taxable income below 2'000'000 CHF can be included. This first step of the data preparation is conducted with the SAS Enterprise Guide. The prepared dataset is then imported in JupyterLab on the Posit Workbench for further preparations.

The main task for the data preparation is the aggregation of the data for the use of the project considering the data privacy restrictions. The data cleaning steps will be discussed in the next subsection. To meet the data privacy restrictions for the registry data, the observations have to be aggregated on a minimum of 10 observations. This is executed with the help of the geocoordinates of the buildings the people live in. Based on the geocoordinates a cell rasterization is applied. This rasterization needs to be dynamic since there are big spatial differences concerning the population distribution in the Canton of Lucerne. Therefore, in a first step the raster size is set on 20 x 20 meters and is filled with observations. Raster cells with 10 or more observations will be excluded for the further rasterization. For raster cells with less than 10 observations the raster size will be enhanced to 30 x 30 meters and filled up again. This process is executed in 10 meter-steps until 2'000 x 2'000 meters. Afterwards, there is just a small amount of observations left, which could not be classified. Furthermore the rasterization is applied over all years in the dataset. Hence the raster cells are sequential over the years. Observations without a successful rasterization are excluded. The following plot shows the distribution of the raster cells size:

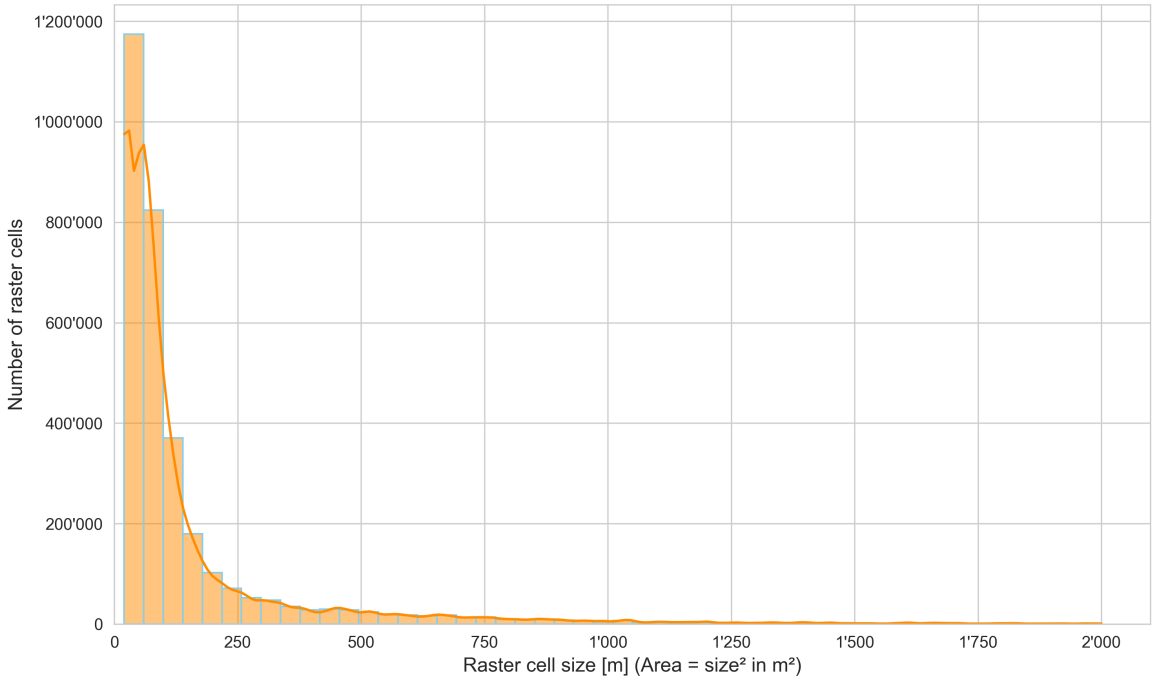

Figure 1: Distribution of the raster sizes.

The distribution of the size of the raster cells is strongly right-skewed, showing that the raster cell size 20 x 20 meters is the biggest group. The dataset before the rasterization contained 3'196'150 observations. Only 23'407 observations (0,7%) could not be assigned to a raster. The rasterization results in the following numbers of raster cells per year:

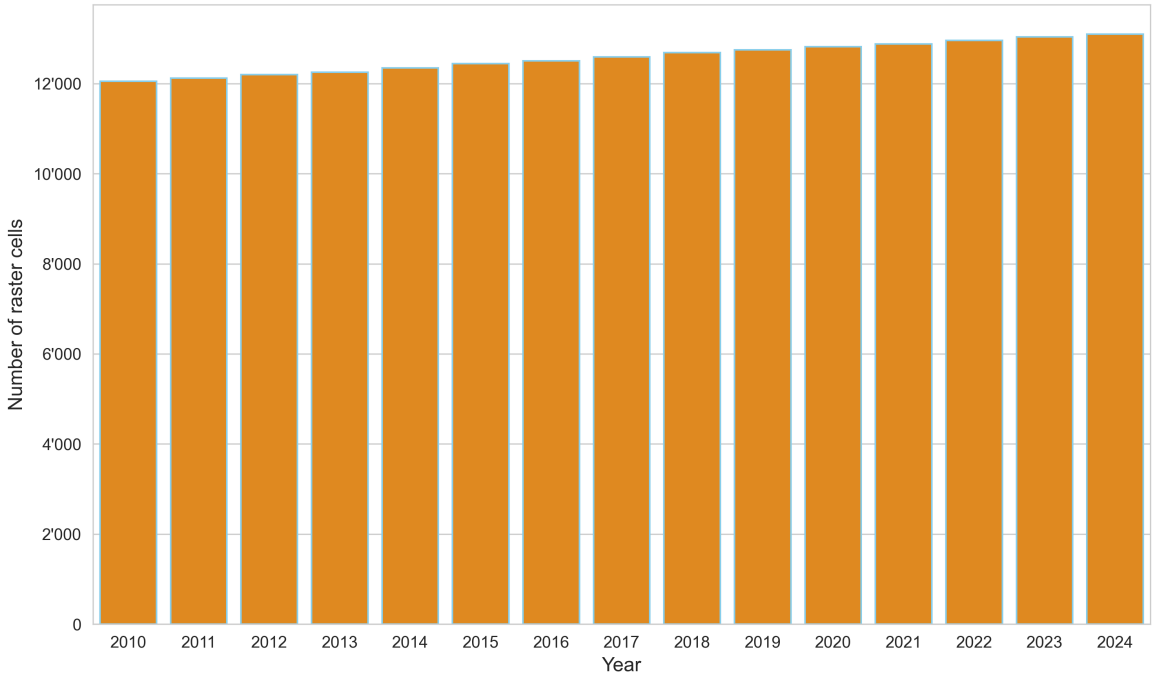

Figure 2: Number of raster cells per year.

The number of raster cells per year slightly increases with each year. This seems plausibel since the population has grown as well and therefore more raster cells with a smaller size are created each year (LUSTAT 2025: 25). To conclude the rasterization, the spatial distribution of the rasters is shown. In the following map, the raster cells with the amount of observations within is illustrated:

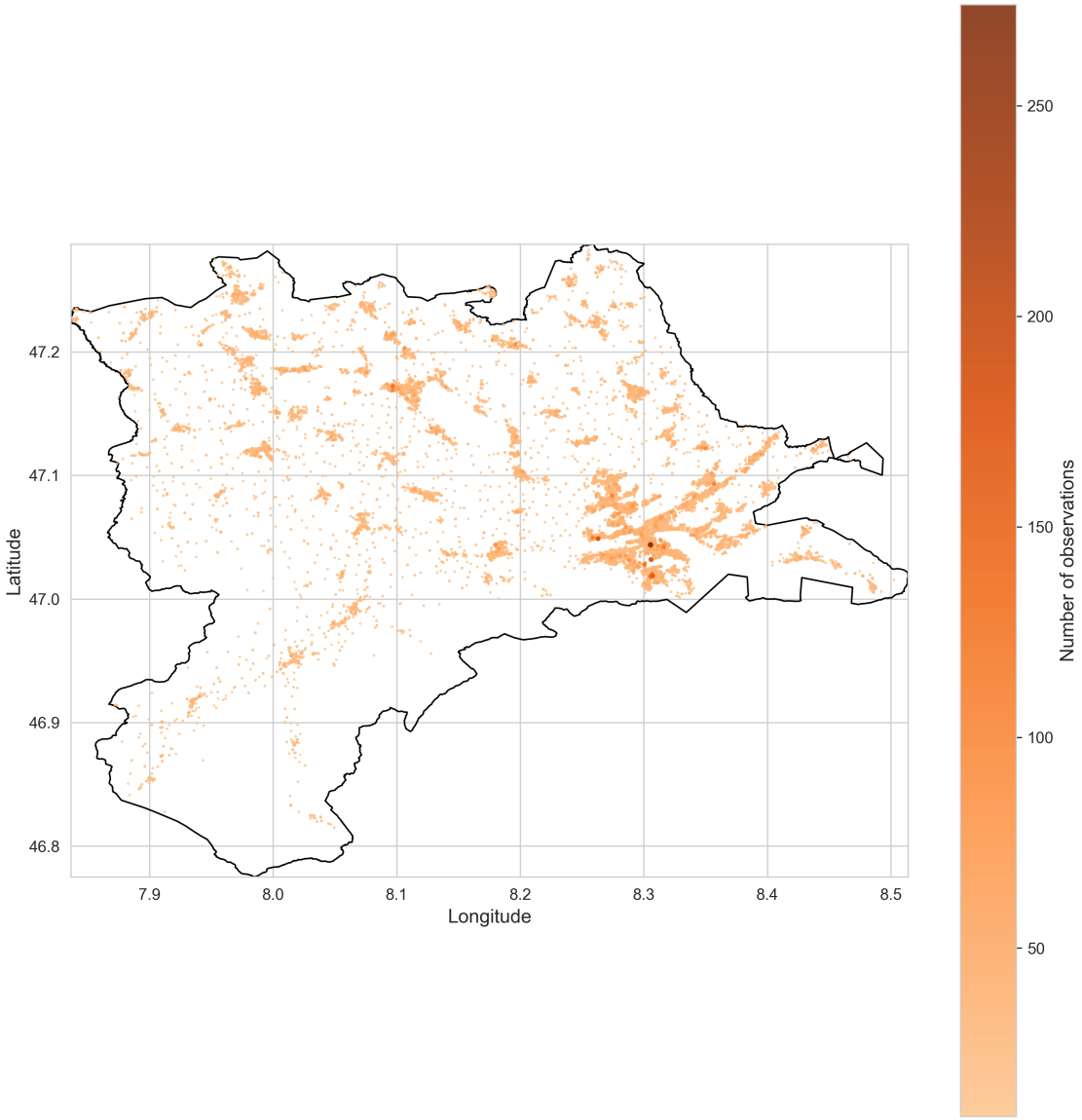

Figure 3: Spatial distribution of the raster cells and number of observations per raster cell.

The distribution of the raster cells and its density seems plausibel. The more urbanized areas around the city of Lucerne are recognizable and the distribution around the lakes (e.g. Lake Sempach) and the main streets in the more rural areas like Entlebuch can be seen as well.

2.2 Data Cleaning

Before the rasterization, the imported data is cleaned with certain measures. The imported raw data contains 3'868'298 observations. Firstly, the observations with a foreign nationality

will be filtered based on their residence permission, since only foreign residents with a C permission (Settled Foreign Nationals) are included in the tax statistics. Foreign residents with other permissions are paying tax at source. This information is not included in the tax statistics. Furthermore, there is a small amount of negative values for the income (7'292 observations). These will be excluded as well. Observations with no information about their tax income (307'829 missings until 2022) are excluded as well. For the years 2023 and 2024 this exclusion cannot be applied, since there is no information available on the tax income for these two years. After the general data cleaning the rasterization is executed. Through the rasterization the observations get a cell identifier (cell id). Based on the cell id the variables in the dataset are aggregated. The following table shows the used variables and how the aggregation is applied:

Table 1: Overview of the used variables including type and aggregation method

Variable Name	**Variable Type**	**Aggregation Method**
Age	numeric	Mean
Civil Status	categorical	Share of Married
Confession	categorical	Share of Protestant Reformed
Income	numeric	Median
Nationality	categorical	Share of Foreigners
Sex	binary	Share of Women
Year	numeric	–

For the income variable the median is used instead of the mean. This should help to mitigate the outliers. Age is operationalized as the mean of each raster cell. The binary and categorical variables are build as the share of one category per variable reaching from 0 to 1. The aggregation of the observations by the cell id leads to the following distributions of the variables:

Table 2: Descriptive statistics of the aggregated variables

Variable	**Count**	**Mean**	**Std.-dev.**	**Min**	**25%**	**Median**	**75%**	**Max**
Age	188'706	41.66	4.28	24.15	38.78	41.64	44.50	60.18
Income	162'578	54'801.84	13'855.85	3'608.00	46'237.25	54'224.00	62'570.75	192'409.00
Share of Foreigners	188'706	0.120	0.149	0.000	0.000	0.071	0.182	1.000
Share of Married	188'706	0.512	0.191	0.000	0.385	0.526	0.643	1.000
Share of Protestant Reformed	188'706	0.101	0.099	0.000	0.000	0.077	0.154	0.933
Share of Women	188'706	0.477	0.092	0.000	0.417	0.474	0.538	0.929

The distribution of the variables will be discussed in detail in the next section Descriptive Data Analysis.

2.3 Data Flow

Since all the steps for preparing the data for the analysis were discussed, the data flow model can be presented as a conclusion of this section:

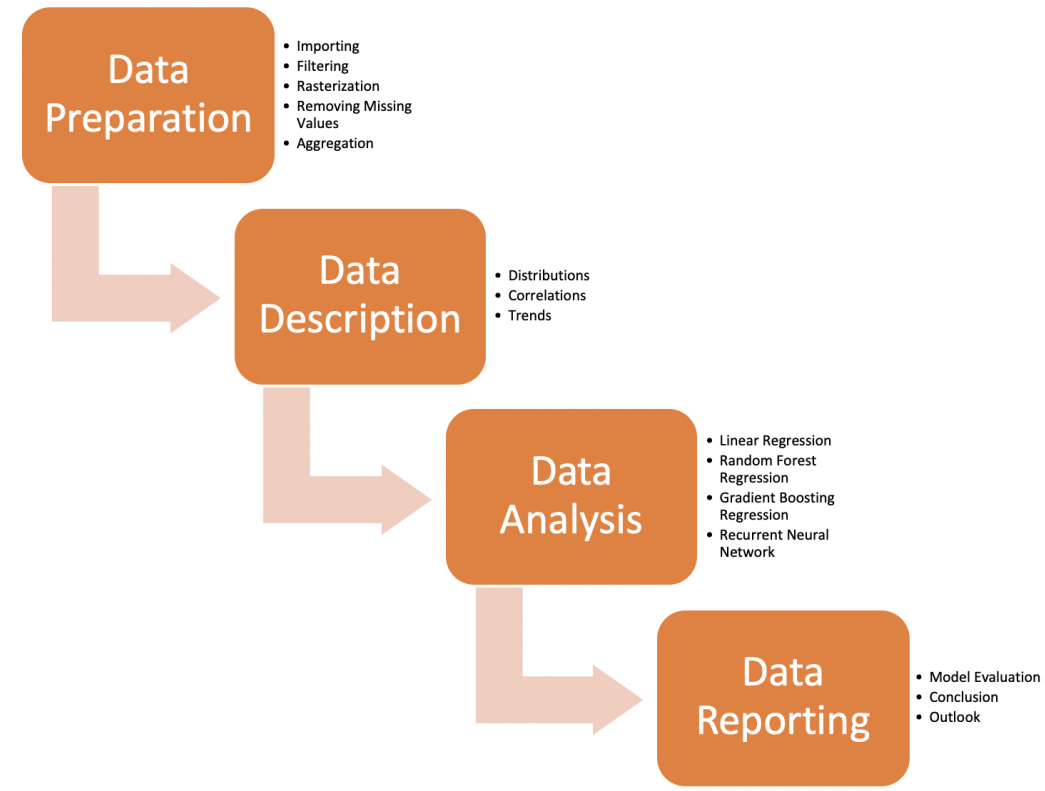

Figure 4: Project Data Flow.

Initially, the data was prepared for the analysis including importing, filtering and removing missing values. Furthermore, a rasterization was applied and the observations were aggregated based on the created raster cells. Secondly, the prepared data is described in form of an descriptive analysis of the distribution of the features and the target variable. The next section will treat this step. Thirdly, the machine learning models are trained, tested and applied to predict the target variable for the years 2023 and 2024. Lastly, the results are reported in form of a model evaluation based on the results of the analysis, conclusions are drawn and an outlook for a possible usage of the tested algorithms for the statistical office is provided.

3 Descriptive Data Analysis

Since the data is now prepared for the analysis, we will take a look at the distribution of the target variable income and the features age, civil status, confession, gender, and nationality (in an aggregated form) before the machine learning analysis is started. In a first step, the target variable income is examined. Secondly, the features are scrutinized. Finally, the relationships between the target variable and the features are discussed.

3.1 Target Variable: Income

The following plot shows the distribution of the target variable income:

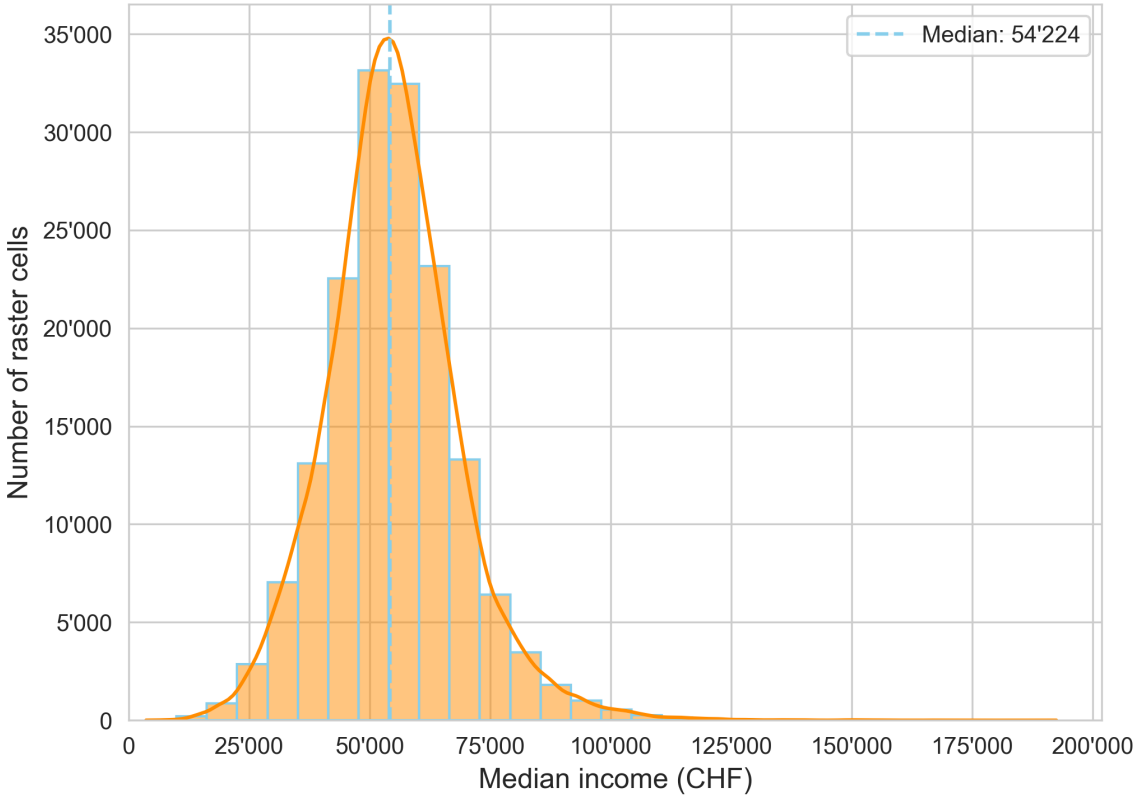

Figure 5: Distribution of the target variable income.

Since there are some extreme outliers with high income values the 100th percentile is excluded. After the exclusion of the 100th percentile the distribution looks as follows:

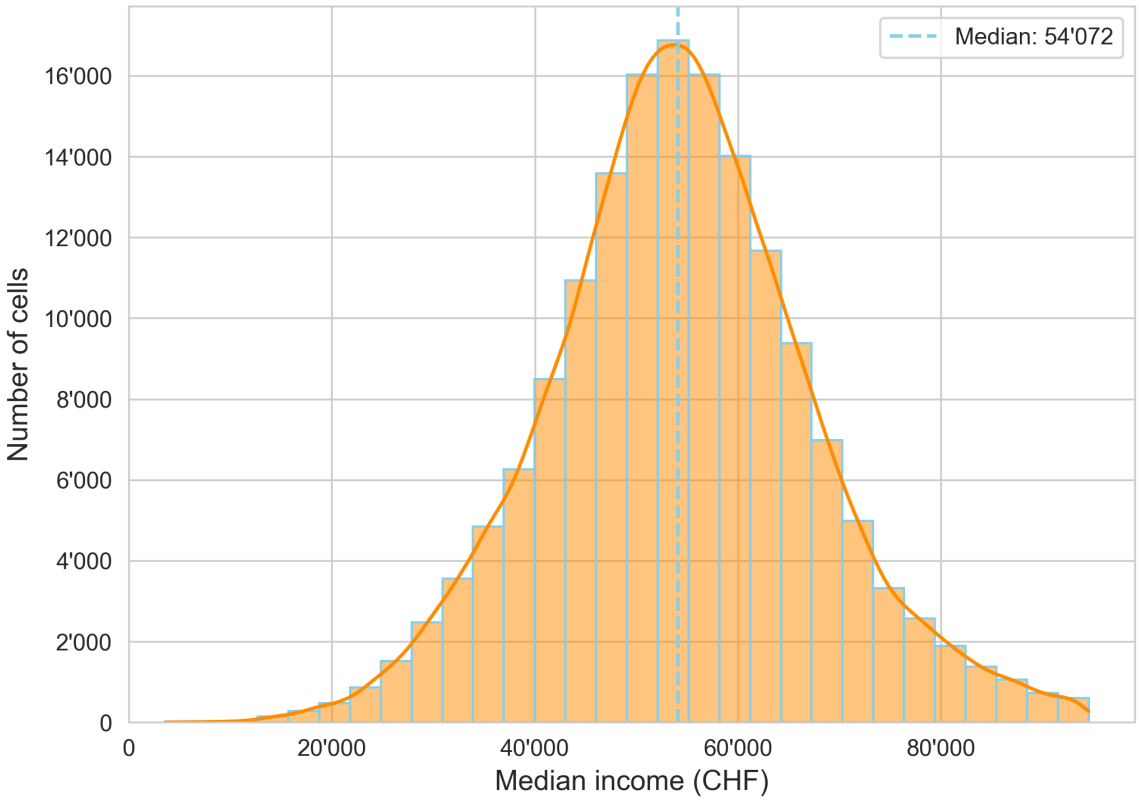

Figure 6: Distribution of the target variable income after excluding the 100th percentile.

With the exclusion of the 100th percentile the distribution shows an almost symmetric form with a median of 54'072 CHF. The aggregation of the income data contributes to the symmetry. Therefore the target variable is ideal to apply regression models on it. For the further understanding of the target variable the development of the income distribution from 2010 is 2022 is plotted:

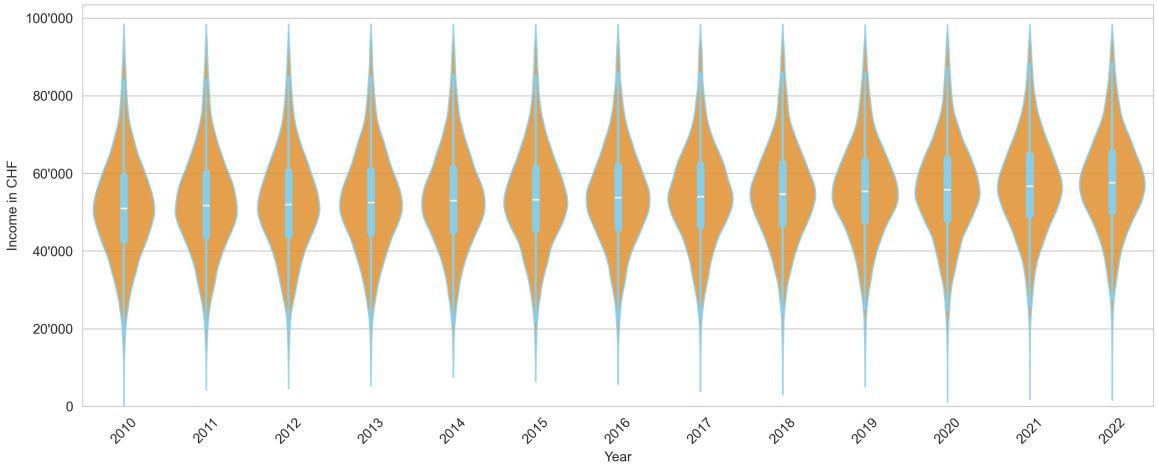

Figure 7: Distribution of the income from 2010 to 2022.

The plot shows that the median of the income is increasing with each year from 2010 to 2022. However, the increase fluctuates between the years. Therefore, the next plot shows the change of the mean of the income compared to the previous year:

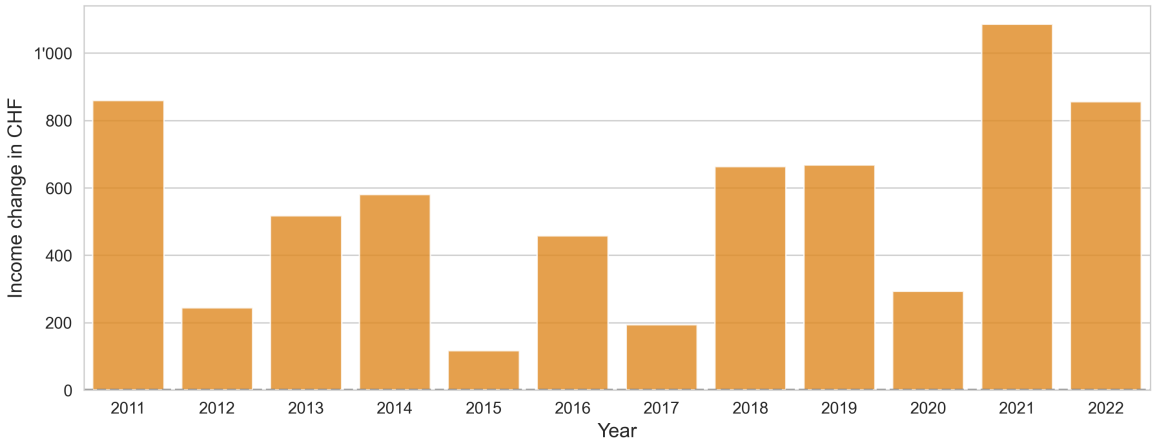

Figure 8: Change in mean income compared to the previous year from 2011 to 2022.

The highest increase compared to the previous year can be observed in 2011, 2021 and 2022. The lowest increase can be observed in 2012, 2015 and 2017.

3.2 Feature Variables

The following plot shows the distribution of the five feature variables:

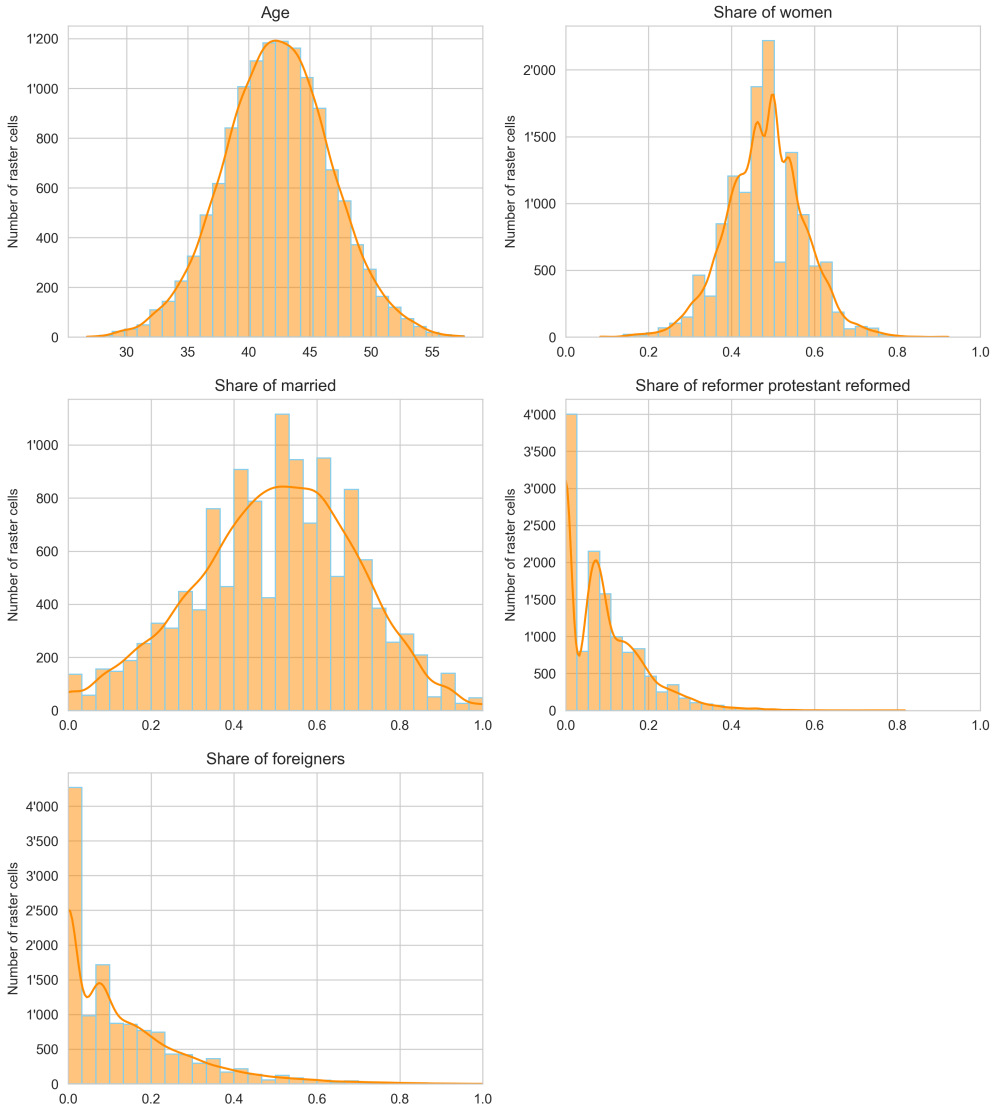

Figure 9: Distribution of the feature variables in 2022.

Age shows a strong symmetric distribution. The distribution of the share of women is also relatively symmetric and shows smaller peaks. The share of married is a bit more heterogenic, slightly left-skewed and has many peaks. The share of protestant reformed and of foreigners are strongly right-skewed with a lot of low values.

3.3 Target and Feature Variables

The following plot shows the relationship between the target variable income and three of the feature variables age, share of women and share of protestant reformers as a linear and a polynomial regression:

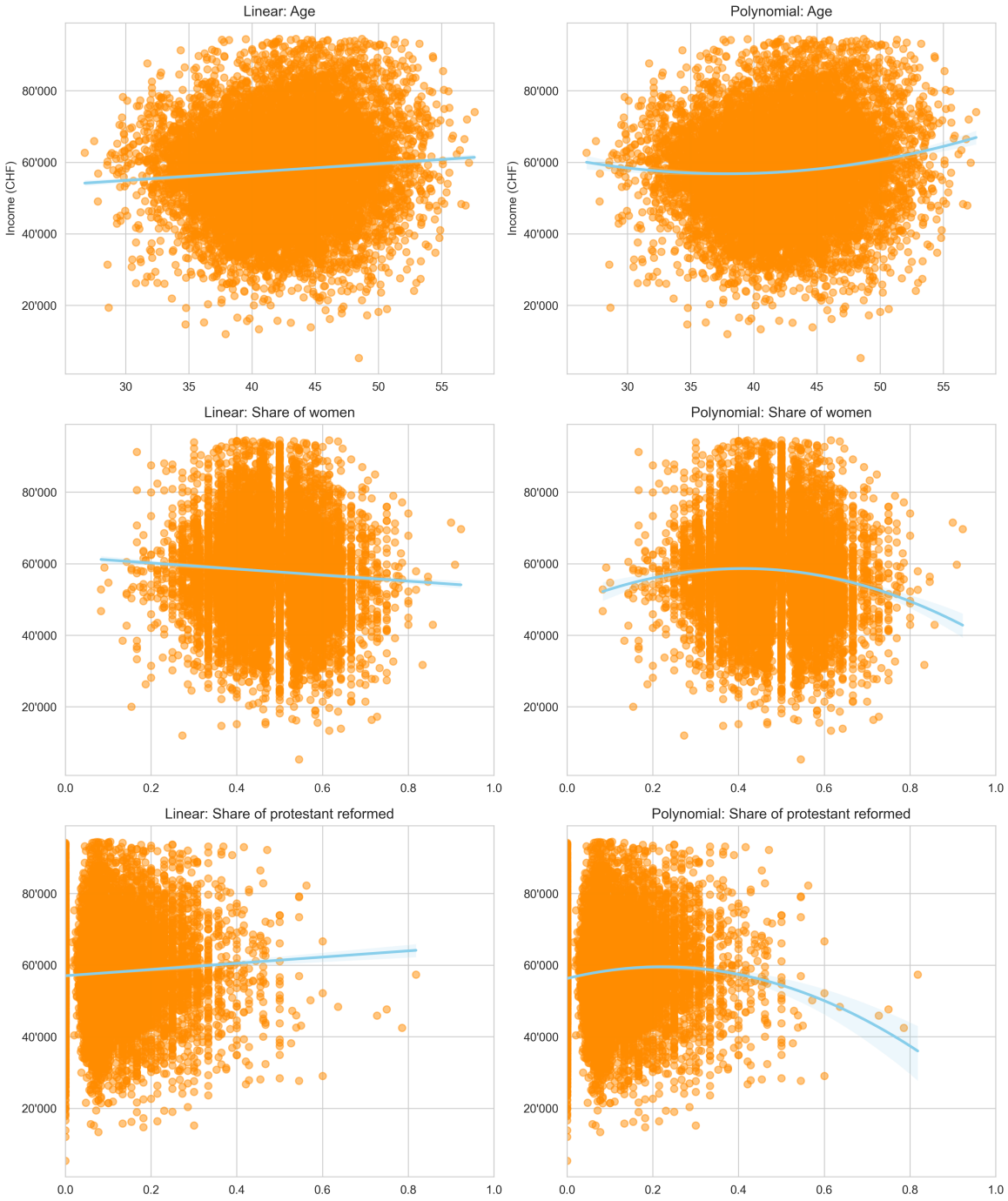

Figure 10: Linear and polynomial relationship between features and target variable income (part 1).

Age shows a weak linear relationship with a slight slope. The polynomial regression shows that the income decreases until around 40 years and increases afterwards. The share of women shows a weak negativ linear relationship with income. The polynomial regression shows that income increases until a share of roughly 0,4 and decreases afterwards. The linear relationship between the share of protestant reformers and income is weakly positive. However, the polynomial regression shows a slight increase at the beginning but decreases with a higher share.

The relationship with the other two features, share of foreigners and share of married, is shown in the following plot:

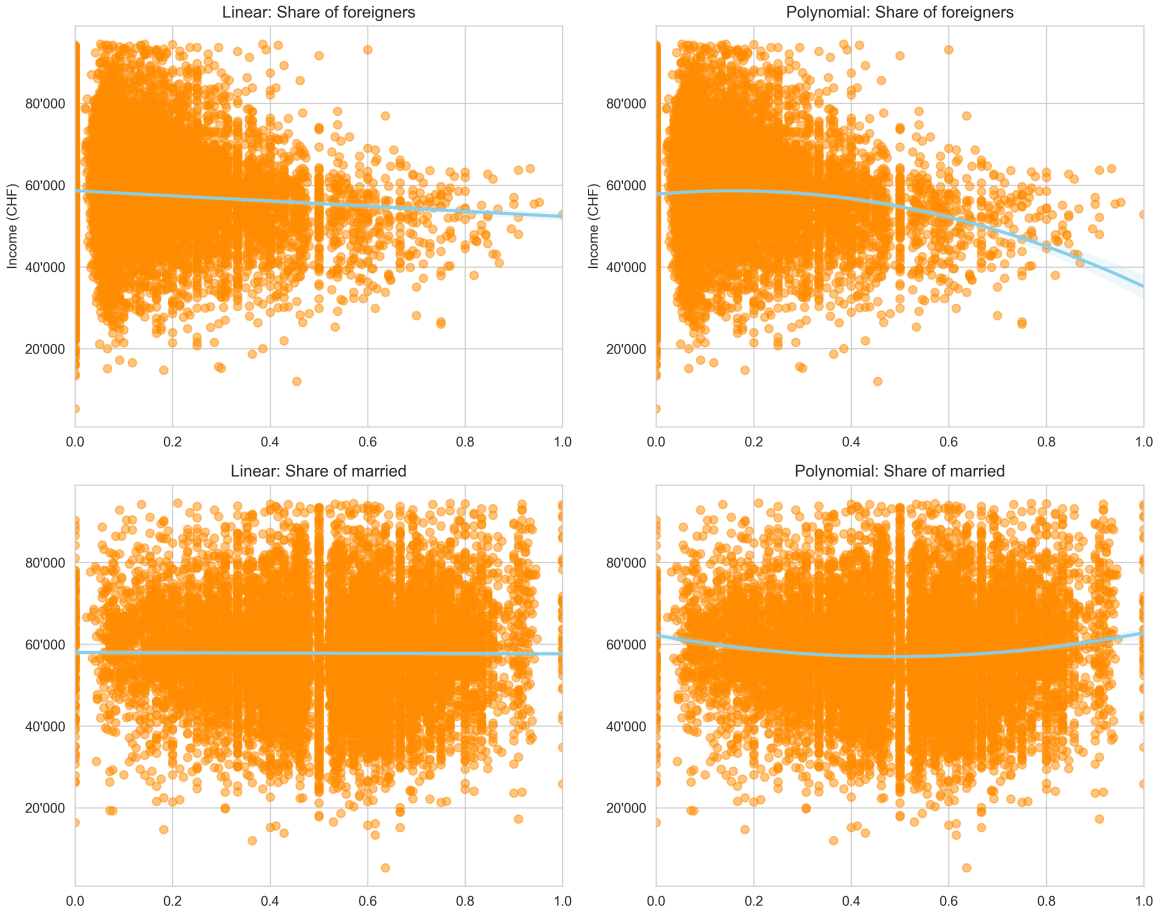

Figure 11: Linear and polynomial relationship between features and target variable income (part 2).

The share of foreigners and the income show a weak negativ linear relationship. The polyno-

mial regression shows a stable development at the beginning. With an increasing share the income decreases exponentially. Therefore these two variables indicate a clear nonlinear relationship. The share of married and the income have no linear relationship at all. The income stays stable for all levels of share. The polynomial regression shows a u-shaped relationship with a higher level of income for the lowest and the highest shares.

4 Machine Learning Analysis

Since the data is now finally prepared and documented for the analysis, the application of different machine learning algorithms can be executed. To compare the advantages and disadvantages of the different machine learning algorithms for income prediction the features are not further transformed. As mentioned earlier, four different algorithms are implemented. Firstly, a linear regression model will be applied. This builds the baseline model. Secondly, a random forest model is implemented. Thirdly, a gradient boosting model is applied. Finally, a recurrent neural network model is built. The models are trained based on the years from 2010 to 2021 and tested on the year 2022. After training and testing, the models are applied to predict the income for the years 2023 and 2024.

4.1 Basic Models

All the models are trained on the sequences from 2010 to 2021 and tested on 2022. The models are trained with no further specific adjustments like hyperparameter tuning. For the RNN-model a SimpleRNN-structure with 64 hidden units and a hyperbolic tangent function (tanh) is applied. There is a basic difference between the RNN-model and the other models. The RNN-model can extract sequential information from the dataset and therefore, only observations which are available in all years from 2010 to 2022 can be used (11'484 observations). The linear and random forest regression models were implemented in Python using the scikit-learn package. The gradient boosting regression model was developed with the xgboost library, and the RNN-model was built using the TensorFlow framework.

4.1.1 Model Quality

As performance metrics the root mean squared error (RMSE) and the mean absolute error (MAE) are calculated. The following plots show the RMSE and the MAE for each model:

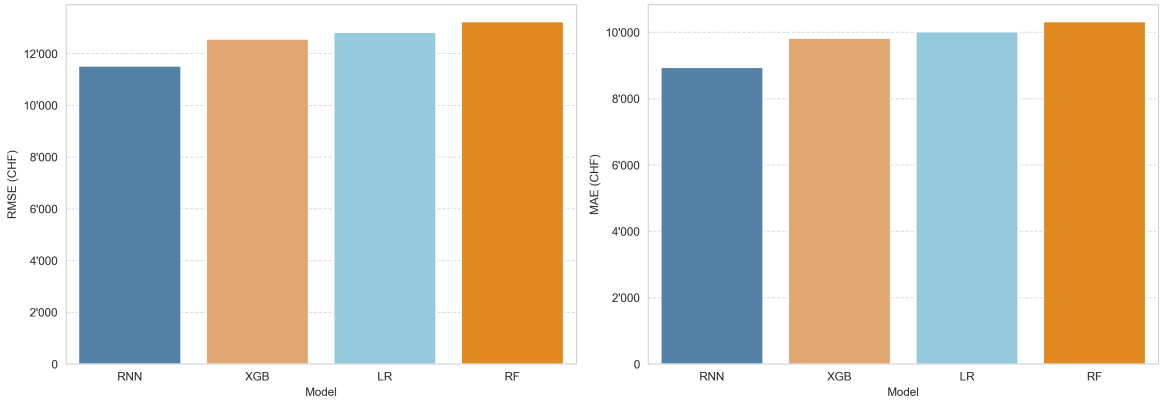

Figure 12: Root mean squared error (RMSE) & mean absolute error (MAE) of the four basic models.

The first plot shows that the RNN-model has, based on the RMSE, the best performance. It is the only model with a RSME lower than 12'000 CHF. The gradient boosting model (XGB) is the second best model. The linear regression model is on the 3th place and the random forest model shows the weakest performance. Looking at the MAE, the results are similar: It seems that even a simple RNN-model outperforms the linear, the random forest and the gradient boosting model. However, the RNN-model has a major advantage: it can learn sequential. The other models cannot extract any information about the sequential development of the income.

4.1.2 Feature & Permutation Importance

To get a better understanding of the functioning of the models the feature coefficients, the feature importance and the permutation importance of the models are discussed. The feature coefficient for a linear regression model indicates the relationship between the features and the target variable. The higher the absolute coefficient the greater the impact of the feature on the target variable (Grömping 2009). The feature importance in random forest and gradient boosting regression models shows how each feature contributes to the accuracy of the model. However, it does not mean, that a feature is uninformative when it has a low value. A low value is just indicating that the feature was not used that much, because another feature encodes the same information (Müller & Guido 2016: 77-78). For the RNN-model the permutation importance is measured. The permutation importance measures the increase in the prediction error of the model after we permute the values of the feature, which breaks the relationship between the feature and the true outcome (Molnar 2022).

The following plot shows the feature coefficients for the linear model and the feature importance for the random forest and gradient boosting model:

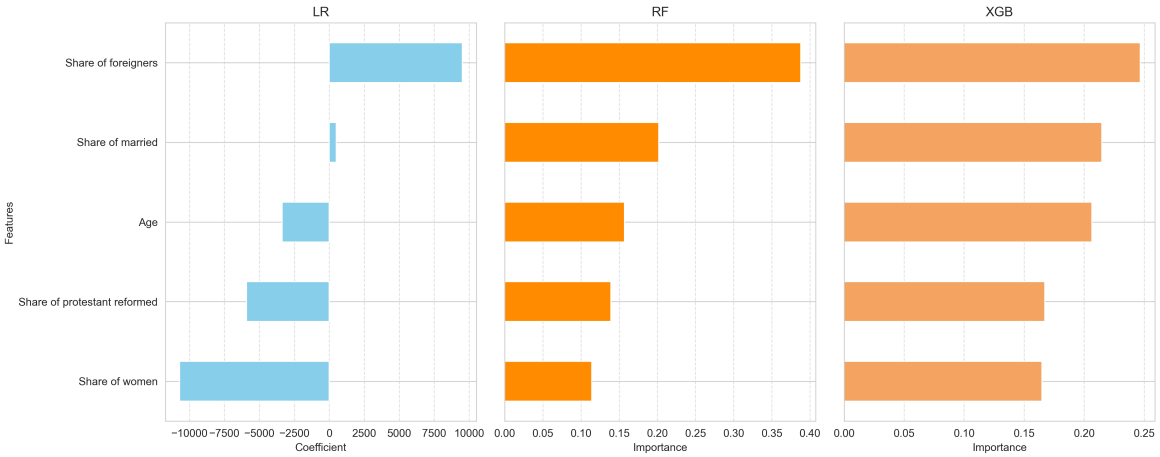

Figure 13: Feature coefficients/importance of the linear, random forest & gradient boosting regression models.

Looking at the linear regression feature coefficients gives the following insights: an increase of the share of foreigners is associated with a higher predicted income. The share of married shows a very small positive effect.The three other features have a negative effect on the income. The strongest negative effect is observed for the share of women. The random forest and the gradient boosting models show pretty similar patterns concerning the feature importance. The share of foreigners is the most important feature, whereas the share of women is the least important. The next plot shows the permutation importance of the RNN-model:

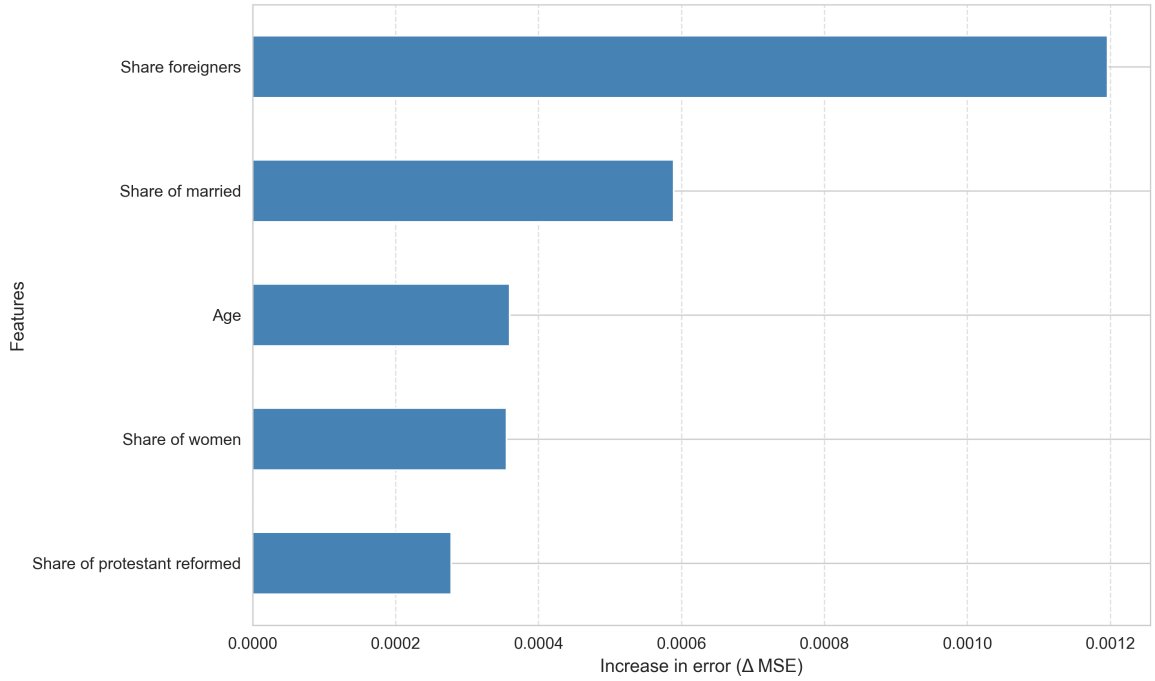

Figure 14: Permutation importance of the RNN-model.

Similar to the random forest and the gradient boosting model, the share of foreigners is the most important feature for the RNN-model. The share of protestant reformed is the least important feature.

4.1.3 Income predictions for 2023 and 2024

Since we have now built up the four models and improved our understanding of the performance of these models, they are applied to predict the income for each raster cell with the help of the five features. The following graph shows the real development of the mean income from 2010 to 2022 and the mean off the estimated income for 2023 and 2024 for each model:

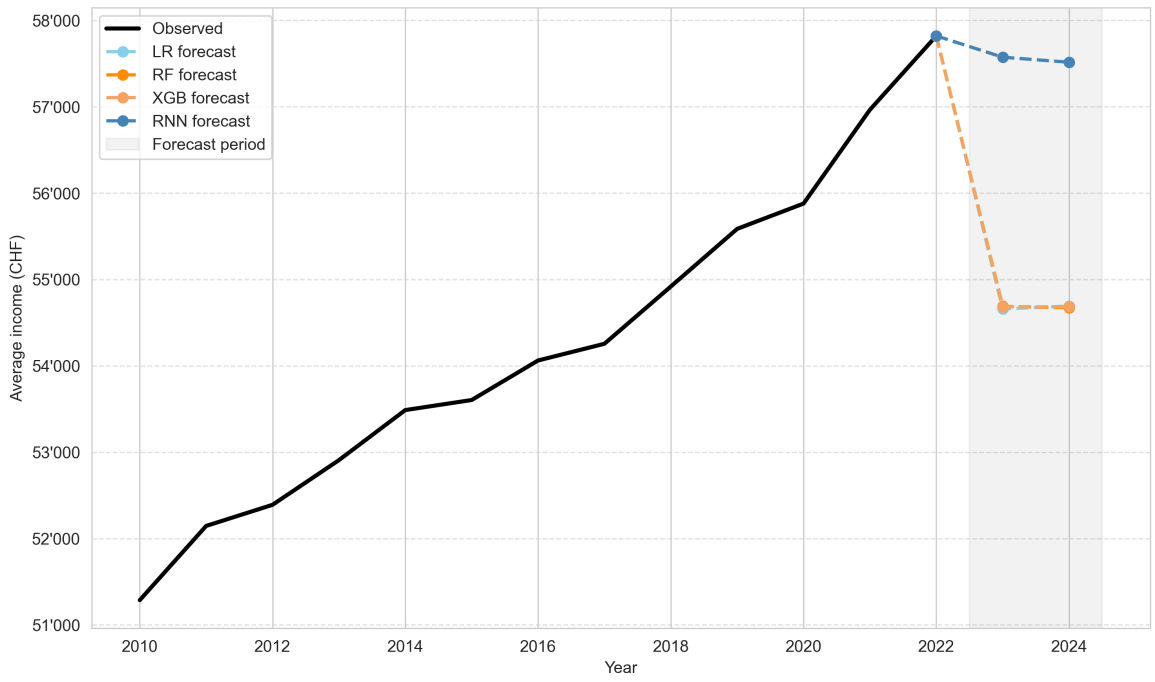

Figure 15: Average of the modeled income by machine learning model.

The graph clearly shows that the RNN-model is the model with the most plausible prediction. However, all models are underestimating the income for the years 2023 and 2024, since an increase in both years can be expected. The other models underestimate the values that can be expected clearly. In the following plot, the uncertainty of the estimations is indicated:

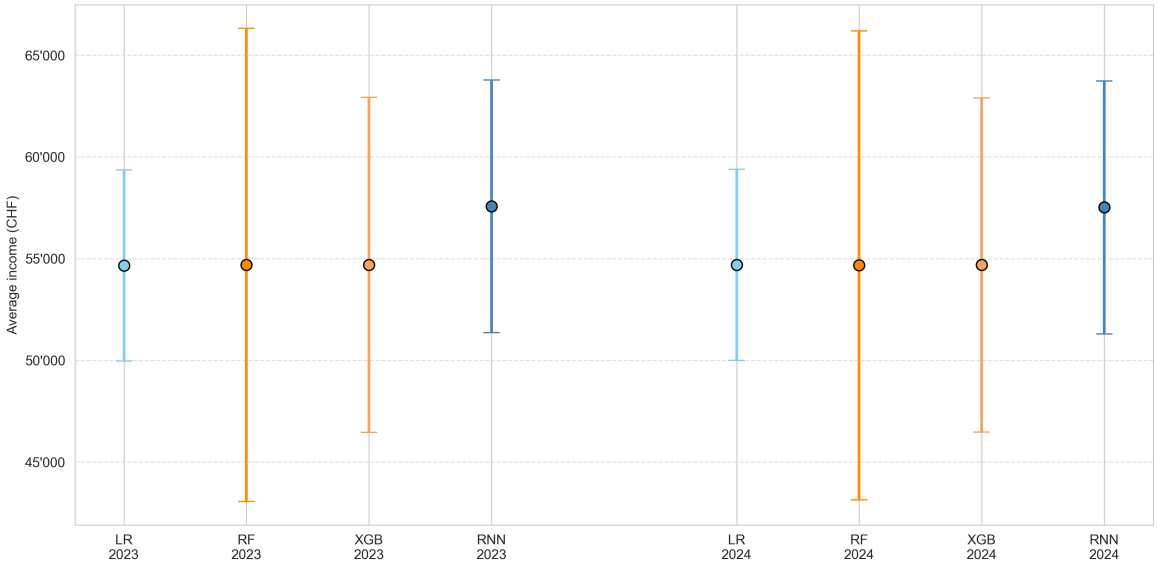

Figure 16: Uncertainty of the modeled income by machine learning model.

There is almost no difference between the linear, random forest and gradient boosting model. This is possibly due to the low variance of the features and the target variable. It seems that the RNN-model already can work with this low variability. Furthermore, the predictions indicate that the RNN-model has a major advantage over the three other models: it can learn sequential from the given data. The other three models do not have any information about the sequential development of the income. The linear regression model has the lowest and the random forest model the highest estimation uncertainty.

4.2 Gradient Boosting Model Improvement

To get a better comparison, the gradient boosting model, which performed the best out of the linear, random forest and gradient boosting models, is improved with the addition of the year as feature and a time-lagged variable. The time-lagged variable takes the value for the four features which are aggregated as share (age is left out) from the previous year (t-1) as additional information for the given year (t). Based on these two new variables two new models are built up: one with just year as additional variable and one with both year and the lagged variables as additional variable. After training and testing the adapted gradient boosting models, they will be compared again with the RNN-model to look if the sequential information helps the model to get an equal or even a better estimation of the income as the RNN-model.

4.2.1 Model Quality

In a first step, the two new gradient boosting models are compared to the basic gradient boosting model without sequential information.

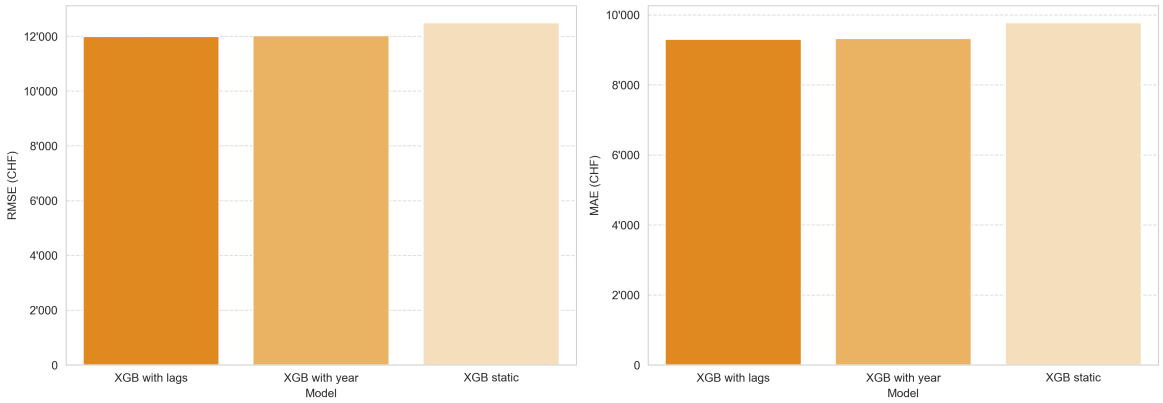

Figure 17: Root mean squared error (RMSE) & mean absolute error (MAE) of the gradient boosting models.

The performance metrics indicate that the model including the year and the lagged variables perform the best, although the difference between the model with just year as additional feature and the model with the lag variables as well, is not that big.

4.2.2 Income predictions for 2023 and 2024

Since the gradient boosting model is now improved with the additional variables, i.e. sequential information, the model with the year and the lagged variables is now applied to predict the income for the years 2023 and 2024:

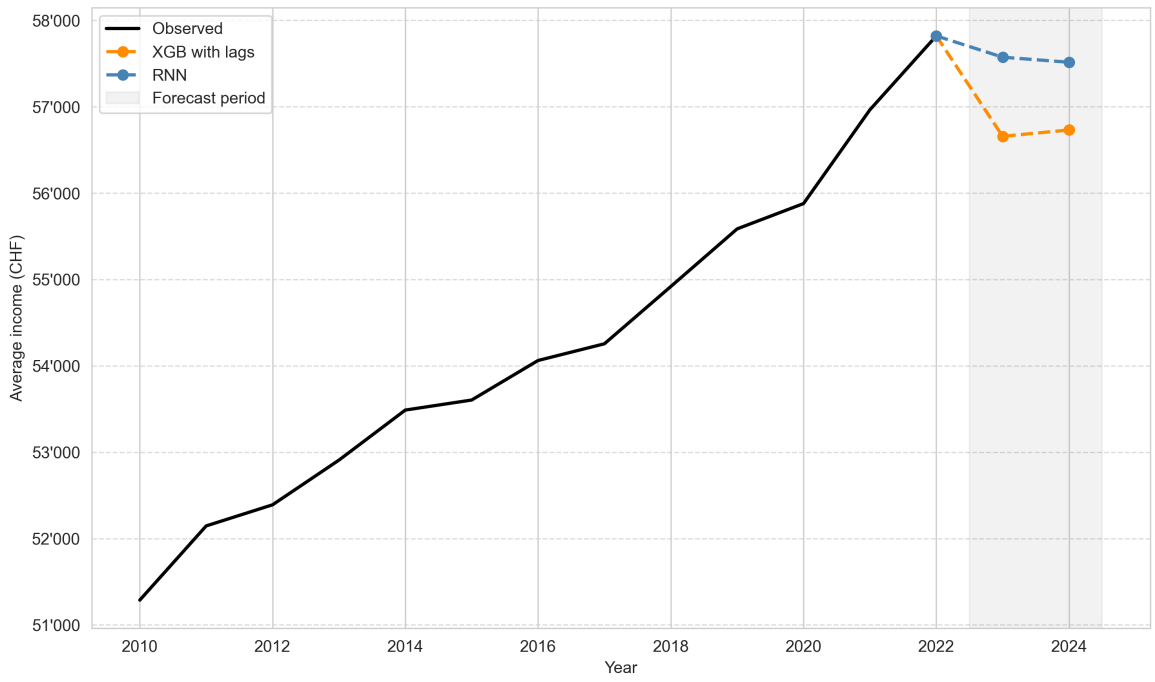

Figure 18: Average of the modeled income by the improved gradient boosting and the RNN-model.

Now with the sequential information, the model performs much better and the estimation for 2023 and 2024 seems more plausible. However, the RNN-Model still shows a better performance. The next plot shows the uncertainty of the estimations of the two models:

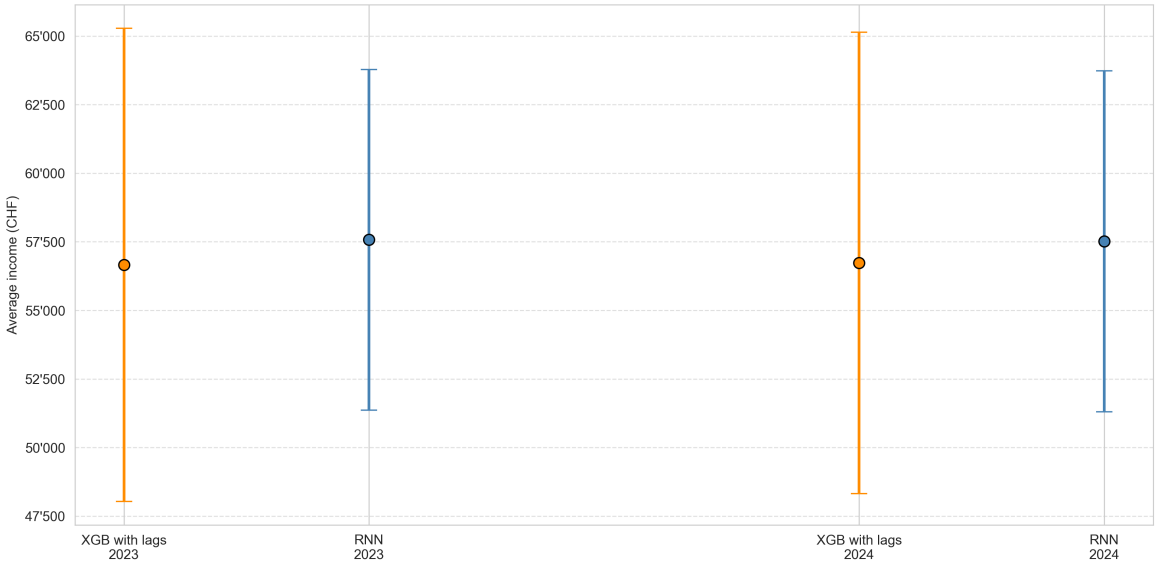

Figure 19: Uncertainty of the modeled income by the improved gradient boosting and the RNN-model.

The uncertainty of the estimation of the gradient boosting model is still higher than the RNN-model. Hence, the inclusion of sequential information for the gradient boosting model helped to increase its performance, but it could not catch up with the performance of the RNN-model with a simple structure. Unfortunately, also the RNN-model cannot provide a plausibel prediction of the income with the given features.

5 Discussion

In the following, the results of the machine learning analysis are discussed. Firstly, the performances of the models are evaluated based on their prediction performance. Furthermore, possible improvements for the different models are presented. Additionally, the handling of the missing income values for the years of 2023 and 2024 are discussed, since this can lead to an estimation bias.

5.1 Model Evaluation

Based on the performance metrics and the plausibility of the predicted income, the RNN-model clearly shows the best performance. The gradient boosting model performed the second best and could catch up a bit to the RNN-model by adding sequential information. Surprisingly, the linear regression model performed better than the random forest model. Possibly, the

used data does not give the random forest model the opportunity to outperform the linear model since the dataset loses information about non-linear relationships through aggregation and also the number of observations is reduced. Hence, the advantage of the random forest algorithm that it can handle very large datasets and non-linearity better than the linear algorithm cannot be benefitted from (Müller & Guido 2018: 88). Probably, for the same reason the gradient boosting algorithm cannot perform as good as possible. However, it is one of the most powerful and widely used algorithm for supervised learning (Müller & Guido 2018: 91). All the models would need further adjustments to enhance their performances. The linear regression model could probably be improved by feature engineering, i.e. by creating interaction terms (e.g. age*gender). Random forest and gradient boosting regression models could be improved for example with hyperparameter tuning (Salman et al. 2024: 77). The RNN model can be enhanced by replacing the SimpleRNN layer with Long Short-Term Memory (LSTM) units (Schmidt 2019: 3). The loss of information, and hence variety, by aggregating the observations on a raster cell level poses also a pivotal challenge to the prediction models. This can be seen if we look for example at the distribution of the income. The 25th percentile lies around 46'000 CHF and the 75th percentile at around 63'000 CHF. The random forest and gradient boosting may perform better with individual data and hence a greater variety. These two algorithms can work with non-linearity and do not need further feature engineering (Müller & Guido 2018). Therefore, with data on the individual level, the random forest and the gradient boosting model can be expected to outperform at least the linear regression model.

A further interesting topic to talk about is the tradeoff between interpretability and performance of a machine learning algorithm. Especially for a statistical office it can also be important that a model is not only good in predicting but also that the model can be understood by the public administration. The following graphic shows the interpretability and the performance of the different machine learning algorithms:

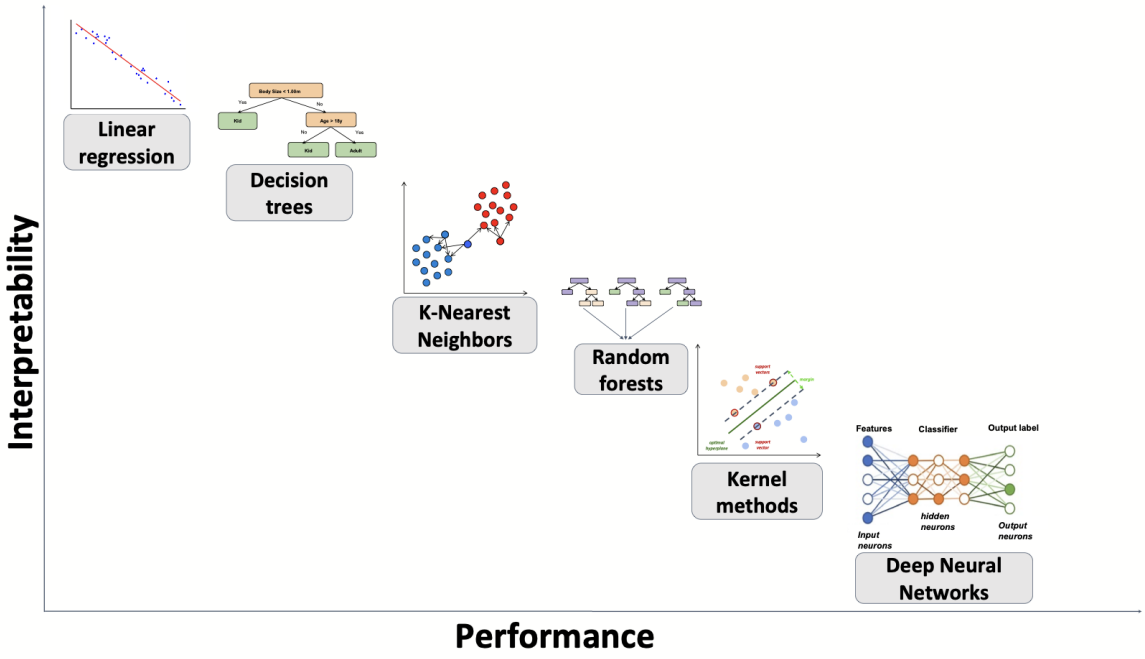

Figure 20: Overview of Machine Learning Algorithms (Source: Badillo et al. 2020: 872).

The linear algorithm provides the best interpretability but shows the weakest performance. The random forest algorithm shows a good balance between interpretability and performance. The gradient boosting algorithm can be put in the same category. The RNN algorithm shows generally the best performance. However, its low interpretability compared to the other algorithms could pose a problem for the understanding for the clients (e.g. public administration). On the other hand, if the methods are laid out clearly, the powerful RNN algorithms could be a good option for prediction models.

5.2 Out-of-distribution detection

In the data preparation for the analysis, the observations with missing income value are excluded in the years 2010 to 2022 for model training and testing. However, this filtering cannot be applied for the years 2023 and 2024 since all income values are missing. Therefore, it has to be checked if this filtering leads to a systematic different distribution of the features in the training and test data (2010 to 2022) and the time period for the income prediction (2023 & 2024). If this is the case then the features for the years 2023 and 2024 would provide an out-of-distribution (OOD) input. An OOD input could lead to an unreliable prediction, since the distribution of the features in the training/testing data would differ significantly from the time period for the prediction. This poses a general problem in machine learning, for example for a autonomous driving system, which needs to detect unusual objects or scenes

it has never seen during training time and hence cannot make a safe decision in this situation (Yang et al. 2024: 1). Therefore, we take a look at the distribution of the features for the year 2022 and 2023 to see if the distribution is systematically different:

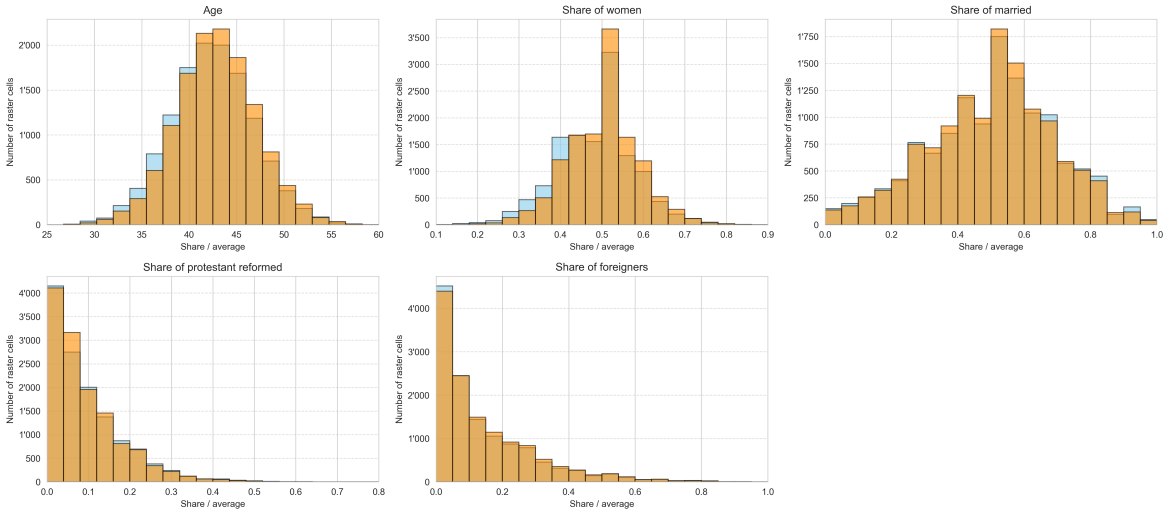

Figure 21: Comparison of feature distributions in 2022 and 2023 for out-of-distribution (OOD) detection.

Generally, the distributions of the features in the years 2022 and 2023 do not show systematic differences between the distributions. Therefore, for this analysis the relatively small differences in the distribution are not expected to contribute to a systematic bias. But it is possible that the small differences in the distribution of the features is attributable to the aggregation of the features. A possible bias through ODD input would have to be checked again when applying the individual data.

6 Conclusion & Outlook

This paper aimed to contribute to the understanding of applying different machine learning algorithms to predictive modeling. More specifically, different models were trained and tested to predict the income with selected socioeconomic factors as features based on register and tax data from the Canton of Lucerne. The machine learning analysis showed that the RNN-model provided the best income prediction. Even a gradient boosting model with sequential information could not keep up with the performance of the RNN-model. However, none of the models could predict a plausible income. Different factors contributed to this lack of performance. Most importantly, the aggregation of the variables on a raster cell level led to a loss of information and variety. Additionally, no further model improvements like feature

engineering or hyperparameter tuning were executed.

Another important issue which was discussed is the accountability, i.e. the interpretability of a model. There seems to be a tradeoff between the performance and the interpretability of the different algorithms. This has taken into account when choosing an appropriate algorithm to build up a prediction model. For example, a random forest model has the advantage that it does not need additional feature engineering in contrast to linear models, like creating interaction variables or normalizing numeric variables. This can save important time with building up a model. However, the communication of the results may be more challenging than for a linear model.

A final evaluation of the possible usage of the tested machine learning algorithms for the statistical office cannot be conducted, since the aggregation of the data leads to a meaningful loss of information, which can have a great impact on the performance of the applied algorithms. Therefore, as a next step, the machine learning algorithms should be trained and tested on the individual level. This could bring more relevant insights for the statistical office. Afterwards, the models could be further improved through measures like feature engineering or hyperparameter tuning. Also the work with other datasets from the statistical office seems promising for making further experiences with machine learning algorithms. In general, the application of machine learning algorithms brings a great potential for the statistical office, for instance in therms of prediction performance but also efficient use of resources. On a more specific level, a qualified statement about the application of these machine learning algorithms at the statistical office to fill up missing data due to a time-lag in data availability, can only be made after the algorithms could be tested on the individual level and after the basic models are further improved.

References

Badillo, Solveig et al. (2020). "An Introduction to Machine Learning". In: *Clinical Pharmacology & Therapeutics* 107.4, pp. 871–885.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. Vol. 1. 2. MIT Press Cambridge.

Grömping, Ulrike (2009). "Variable importance assessment in regression: linear regression versus random forest". In: *The American Statistician* 63.4, pp. 308–319.

LUSTAT (2025). *Jahrbuch Kanton Luzern*. LUSTAT Statistik Luzern.

Molnar, Christoph (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd. Independently Published.

Müeller, Andreas C and Sarah Guido (2018). *Machine learning avec Python*. First interactive.

Salman, Hasan Ahmed, Ali Kalakech, and Amani Steiti (2024). "Random forest algorithm overview". In: *Babylonian Journal of Machine Learning* 2024, pp. 69–79.

Schmidt, Robin M (2019). "Recurrent neural networks (rnns): A gentle introduction and overview". In: *arXiv preprint*.

Yang, Jingkang et al. (2024). "Generalized out-of-distribution detection: A survey". In: *International Journal of Computer Vision* 132.12, pp. 5635–5662.

Statement

"Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen."

Date: 9.6.2025 Signature: 

lustat
statistik luzern

Vertrag

über die Bearbeitung von Daten aus
der Registerstatistik

zwischen

LUSTAT Statistik Luzern
Burgerstrasse 22
6002 Luzern

und

David von Holzen
Ziegeleiweg 1
6048 Horw

vom

9. Mai 2025

1. Gegenstand dieser Vereinbarung ist die Nutzung der unter Ziffer 2 genannten Daten durch den unter Ziffer 4 genannten Mitarbeiter von LUSTAT Statistik Luzern im Rahmen seiner Teilnahme an der externen Weiterbildung (CAS Applied Data Science der Universität Bern). Die Daten dürfen ausschließlich für die Abschlussarbeit im Rahmen der genannten externen Weiterbildung genutzt werden.
2. Die verwendeten Daten stammen aus der Registerstatistik des Kantons Luzern, welche Angaben aus dem kantonalen Einwohner- und Gebäuderegister sowie der Steuerstatistik verknüpft und anonymisierte Angaben gemäss Anhang für den Kanton Luzern umfasst. Für die Analyse werden die personenbezogenen Daten auf Gebietszellen mit min. 10 Personen aggregiert.
3. Der Datennutzer verpflichtet sich, die Daten nur für den unter Ziffer 1 vereinbarten Zweck zu verwenden und diese ohne ausdrückliche schriftliche Zustimmung des Datengebers nicht an Dritte weiterzugeben.
4. Die unter Ziffer 2 genannten Daten dürfen seitens des Datennutzers nur von folgenden Personen eingesehen und bearbeitet werden:
   - David von Holzen (wissenschaftlicher Mitarbeiter)
5. Diese Vereinbarung gilt für die Dauer der Teilnahme an der genannten Weiterbildung und endet automatisch nach deren Abschluss, sofern keine anderslautende schriftliche Vereinbarung getroffen wird.

LUSTAT Statistik Luzern

Tim Hagmann
Direktor

LUSTAT Statistik Luzern

Roberto Frisullo
Mitglied der Geschäftsleitung

Datennutzer

David von Holzen
Wissenschaftlicher Mitarbeiter

Anhang: Umfang des Datensatzes

1. Grundgesamtheit: Personen im Alter von 18 bis 64 Jahren mit einem Erwerbseinkommen unter 2'000'000 Fr..
2. Datenjahre: 2010 bis 2024
3. Variablenliste:

Variable	**Beschreibung**
jahr	Registerstatistik / Fachliches Referenzjahr
rst_alter	Registerstatistik / Vollendetes Altersjahr am Jahresende
ewr_sex	EWR / Geschlecht (Code)
ewr_zivilstand	EWR / Zivilstand (Code)
ewr_konfession	EWR / Konfession (Code)
ewr_natldnr	EWR / Nationalität (BFS-Nr.)
ewr_respermit	EWR / Ausländerkategorie (Code)
ewr_hhart	EWR / Haushaltsart (Code)
gwr_gkode	GWR (Gebäude) / E-Gebäudekoordinate
gwr_gkodn	GWR (Gebäude) / N-Gebäudekoordinate
sn_erwerb_p1_p2	Steuerstatistik NP / Einkommen aus Erwerbstätigkeit