Predicting Household Welfare Outcomes Using Observable Socio-Economic Characteristics

Daniël T.J. Molenaar
9879439

Master thesis
Credits: 18 EC

Master *Applied Data Science*

First supervisor
Mahdi Shafiee Kamalabad

Second supervisor
Samin Nikkhah Bahrami

Informatics Institute
Faculty of Science
Utrecht University
Princetonplein 5
3584 CC Utrecht

Semester 2, 2024-2025

Abstract

This thesis explores the predictive capacity of machine learning models to estimate household welfare outcomes using socio-economic characteristics within the context of Index-Based Livestock Insurance (IBLI). IBLI aims to mitigate the economic shocks of climate-induced livestock loss by offering insurance based on satellite data. While previous studies have shown the heterogeneous welfare effects of IBLI, this thesis investigates whether machine learning methods are effective at uncovering heterogeneous welfare effects across household subgroups and can show the most important household characteristics.

Four models have been evaluated: Lasso Regression, TabTransformers, Generalized Random Forest, and Bayesian Ridge Regression, using cattle and goat datasets. These datasets originate from a dataset with information about herders in southern Ethiopia who make IBLI purchase decisions for their herd. Results show that all models performed better on the goat dataset, with TabTransformers outperforming others in terms of predictive power. Despite modest overall performance, key features such as the settlement status of households and trust in village insurance promoters consistently emerged as key predictors. These findings highlight both the capabilities and limitations of applying machine learning in these contexts.

Keywords: Index-Based Livestock Insurance, Machine Learning, Welfare prediction, TabTransformers, Generalized Random Forest, Lasso, Bayesian Ridge

Contents

Chapter 1

Introduction

Natural disasters continue to be one of the most devastating forces affecting livelihoods across the globe. In the arid and semi-arid regions of Sub-Saharan Africa, droughts are not occasional anomalies but recurring threats. For communities whose income and food security depend on livestock, the stakes are high. Despite this, access to conventional insurance markets remains limited.

Index-Based Livestock Insurance (IBLI) has emerged as a promising financial innovation to help decrease the impact of climate shocks on vulnerable households. As can be seen on the IBLI website(International Livestock Research Institute (ILRI), 2025), IBLI triggers payouts using satellite-based vegetation indices such as the Normalized Difference Vegetation Index (NDVI), which is compared with historical averages to identify drought conditions, rather than relying on costly and time-consuming evaluations of individual farm or herd losses.

This insurance structure offers scalability and a potential safety net for communities living in volatile climates(Chantarat et al., 2017). "By providing an ex-ante mechanism for smoothing income in response to climatic shocks, IBLI aims to reduce reliance on harmful coping strategies such as distress sales or reduced food consumption and to promote productive investments and long-term resilience" (Barrett et al., 2019).

However, the real-world impact of such products is not uniform. IBLI as it stands does not benefit every household equally(Harrison et al., 2020). Different households experience different levels of welfare beneftis: the impact of IBLI is heterogeneous. Different socio-economic factors play an important role in the adoption of IBLI(Takahashi et al., 2016). The question remains: What are the subgroups of people who gain welfare from this product? This thesis aims to investigate the heterogeneous welfare effects of IBLI in Ethiopia by applying machine learning algorithms on two different datasets: a goat dataset and a cattle dataset. These two datasets originate from a singular dataset, containing data collected in southern Ethiopia. The data in the used dataset ranges from socio-economic information to demographic information. The primary goal is to develop, train and fine-tune a set of predictive models to identify the approach that yields the highest performance, and identify the characteristics that predict welfare outcomes. By comparing these models using various metrics, this research aims to determine the most effective method for analysing the IBLI dataset and delivering reliable

predictions.

While several studies have already evaluated the impact of IBLI on welfare indicators, less attention has been given to the heterogeniety in these outcomes across different households. Understanding which household characteristics are most predictive of welfare outcomes can help with the design of IBLI targeting, IBLI as a product, targeting IBLI to certain households, and the design of risk mitigation strategies. In this context, machine learning (ML) offers a valuable set of tools. ML can help uncover complex, non-linear relationships between welfare and household characteristics that might not be captured by traditional econometric methods.

This thesis will explore whether machine learning models can improve our ability to identify and predict welfare outcomes among households in Ethiopia. By comparing a variety of algorithms, this research aims to assess both the predictive performance and the interpretability of these models in the context of welfare analysis. The question guiding this research is:

How effective are machine learning methods in predicting household welfare outcomes using observable socio-economic characteristics?

To support this research question, a subquestion has been drafted:

What are the most important household characteristics used to predict welfare on the IBLI dataset?

Chapter 2

Theoretical Framework

This theoretical framework will outline the key concepts, mechanisms, and modelling approaches that underlie the use of IBLI and its interventions. It will draft the context in which this thesis will operate by using IBLI as a tool for climate risk mitigation and evaluating the role of predictive and causal modelling in enhancing the effectiveness of IBLI.

2.1 Index-Based Livestock Insurance (IBLI)

Index-Based Livestock Insurance (IBLI) is a product designed to protect households in arid and semi-arid regions against the effects of drought on livestock mortality. Unlike traditional insurance, IBLI pays benefits based on objective and remotely detected indices related to vegetation or rainfall, which serve as proxies to calculate the risks of mortality from livestock (Chantarat et al., 2017).

IBLI was developed as a response to fundamental structural barriers in providing insurance to pastoralist communities. Traditional insurance models are often not feasible in these regions due to high transaction and verification costs, difficulties in observing individual losses, and the potential for moral hazards and adverse selection, according to Chantarat et al., 2008. These challenges are especially pronounced in low-income and remote areas, where access to formal financial institutions is limited and risk-sharing mechanisms are weak(Barrett et al., 2019). IBLI addresses these issues by relying on satellite-based indices, such as the Normalised Difference Vegetation Index (NDVI), to trigger payouts, thereby reducing the need for costly and subjective loss verification processes and providing a scalable and transparent solution for managing climate-related risk (N. Jensen et al., 2017).

IBLI has been extensively piloted in Ethiopia and Kenya, where communities are particularly vulnerable to recurrent droughts. Evaluations have shown promising impacts on household resilience and asset protection, especially in the face of climate shocks. However, the effectiveness of IBLI depends heavily on accurate calibration of the index, as well as household understanding and trust in IBLI itself (Barrett et al., 2019; Harrison et al., 2020).

Other interventions such as subsidies and information campaigns have been

tested to increase the adoption and understanding of IBLI. However, concerns remain about the heterogeneity of its welfare effects, especially among the poorest households (Barrett et al., 2019; Harrison et al., 2020). This raises questions about targeting and personalisation, with which machine learning models could offer a helping hand.

2.2 Behavioral factors and heterogeneous adoption of IBLI

While IBLI has shown promising results as a climate risk mitigation tool, its adoption and impact vary greatly between households. These patterns indicate a complex relation between behavioural factors, economic constraints, and the insurance product itself.

2.2.1 Risk perception

One of the main limitations to the wider adoption of IBLI is the so-called *basis risk*, which refers to the discrepancy between the actual losses experienced by a household and the payouts triggered by the index. This risk arises because IBLI relies on remote indicators such as the NVDI to check forage availability across a broad area. However, in spatially heterogeneous environments, two households within the same area may experience very different drought impacts. When the index fails to reflect the actual conditions experienced by a household, it could trigger a payout when no loss occurs or fail to trigger one when losses are there, resulting in a decrease of the trust in the insurance product. As Chantarat et al., 2017 argue, the presence of this basis risk can significantly reduce the demand for index insurance. This has been supported by field evidence from Kenya and Ethiopia, where households report dissatisfaction when payouts do not align with the perceived hardship (N. Jensen et al., 2017). Reducing basis risk remains one of the technical challenges in scaling up IBLI.

2.2.2 Socio-economic factors

Socio-economic factors also play an important role in the adoption and welfare implications of IBLI. Education, general wealth, and access to financial infrastructures all affect the likelihood of the household purchasing IBLI. "Education and wealth positively influence the likelihood of purchasing IBLI" (Takahashi et al., 2016). These characteristics often influence both the understanding of the product and the capacity to pay the premiums.

2.2.3 Behavioural responses and heterogeneous impacts

Having insurance seems to affect household behaviour. According to the article in N. D. Jensen et al., 2024, the coverage of IBLI increases investments in intensifying livestock production, specifically in animal health. This can also be seen in the

findings of the article by Karlan et al., 2014. This research shows that farmers, who have received the rainfall index insurance, have increased their investments by 13%, compared to relatively small changes for farmers who have not received the insurance. In addition to this, households adapt more efficient market strategies. Households "sell more livestock during nondrought seasons when prices are high and avoid 'distress selling' during droughts when prices are low" (N. D. Jensen et al., 2024). These findings show a behavioural transformation that underscores not only the potential positive effect of IBLI but could also serve as a catalyst for the local economy.

Despite these promising outcomes, IBLI in its current form does not benefit every household equally. "IBLI has heterogeneous impacts, with the most significant benefits observed among households with moderate herd sizes and previous exposure to insurance." (N. D. Jensen et al., 2015). For poorer households, financial constraints remain a barrier.

2.3 Data and models

This section highlights the structure of the dataset used in the analysis, followed by a brief overview of the models used to estimate heterogeneous effects and identify key predictors of welfare outcomes in the context of IBLI.

The dataset contains detailed household-level information collected from 2416 Ethiopian pastoralist households across three survey rounds, with 6 months in between rounds. Each household may have made up to six insurance decisions, covering purchase and non-purchase IBLI decisions for cattle and goats over three time periods. The dataset contains both purchase and non-purchase decisions and links these to two constructed welfare measures. The first welfare measure is a continuous difference-based measure that compares actual and advised insurance decisions, while the second is a normalised welfare ratio that accounts for the differences in the size of the herd.

In addition to the outcome variable, the dataset contains a rich array of covariates. These covariates include demographic attributes like age and gender, economic indicators like herd size and income sources, behavioural variables like the trust in IBLI information provides, and contextual information like location and owning a phone. Some variables are or have been made time-invariant, while others vary across rounds.

The presence of missing data and the structure of certain variables required careful preprocessing, including imputation, transformation, and encoding strategies. These steps have been explained in Section 3.1. The steps ensured that the final input was suitable for the machine learning models.

2.3.1 Model architectures

Understanding the effects of insurance uptake on household behaviour requires models that can both predict outcomes accurately and identify heterogeneity in treatment effects. For this research, a combination of statistical learning tools are

used that allow for parsimonious variable selection (one that uses relatively few independent variables to obtain a good fit to the data) in the shape of a Lasso regression, nonlinear feature interaction modelling in the shape of TabTransformer, a causal heterogeneity estimation in the shape of a Generalized Random Forest, and a bayesian model in the shape of the Bayesian Ridge model.

2.3.2 Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) is a regularisation method that performs both variable selection and shrinkage, making it especially useful in high-dimensional settings (Tibshirani, 1996). The Lasso helps to identify the most relevant predictors while preventing overfitting by penalising the absolute size of coefficients. By doing this, Lasso forces some of them to become exactly zero, thereby effectively selecting a simpler model.

This property makes Lasso a particularly useful model when there is a need to identify the most important predictors from a large pool. In this thesis, Lasso is used not only to predict welfare outcomes but also to uncover which socio-economic characteristics are most strongly associated with those outcomes. The trade-off between model complexity and its generalisation performance is controlled by the tuning parameterλ.

The equation below shows the formula Lasso uses to minimize the sum of squared errors between the predicted and actual values, including a penalty term which encourages sparsity in the coefficients.

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.1}$$

2.3.3 TabTransformers

TabTransformers (Huang et al., 2020) is a deep learning architecture designed specifically for tabular data, combining the power of transformers with categorical embeddings. Traditional neural networks often struggle with categorical features unless they are heavily processed. Tabtransformers addresses this limitation by learning rich embeddings and capturing complex feature interactions. In this application, the model is particularly suited to capture subtle interactions between insurance coverage and household characteristics that can influence behavioural outcomes.

The core innovation lies in the use of transformer layers which leverage multi-head self-attention to dynamically model complex interactions between features. This is a particularly useful feature when dealing with tabular data settings where relationships among categorical variables and continuous variables can be non-linear and subtle.

In figure 2.1 , a schematic view of the workings of a TabTransformer can be found. The figure illustrates the architecture of TabTransformers, which processes tabular data by embedding categorical features and passing them through a stack

of transformer layers. These layers use multi-head attention to capture interactions between categories. Continuous features are normalised separately and later concatenated with the transformer output. This combined representation is then passed through a feedforward neural network to make final predictions. The design enables the model to learn both feature importance and complex relationships within the data.

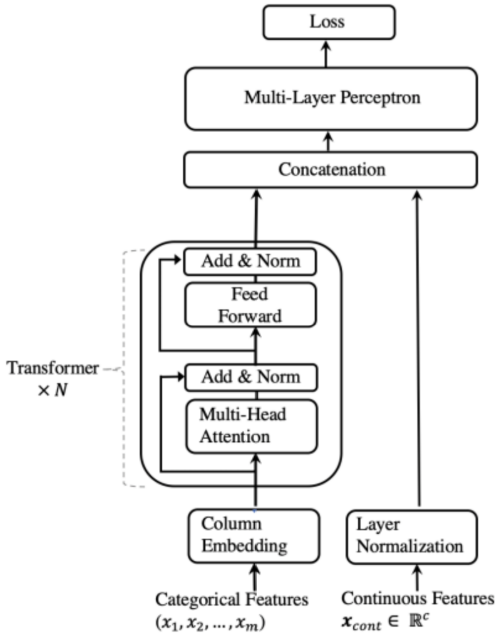

Figure 2.1: The architecture of TabTransformers, from Amazon Web Services, 2024

Overall, the design of TabTransformers enables the extraction of complex patterns and feature importances from tabular datasets without extensive manual feature engineering. This makes them a strong alternative to more classic methods like gradient boosting or random forest models.

2.3.4 Generalized Random Forest

The Generalized Random Forest (GRF) framework (Athey et al., 2019) extends traditional random forests to estimate conditional average treatment effects (CATE). This approach enables us to identify the impact of insurance differences across households. Next to this, it will also provide insights into heterogeneous impacts that are central to the policy debate.

The key innovation in GRF is the use of honest estimations. This is a methodological improvement that splits the sample into two subsets: one used for determining the tree structure and another used for estimating the treatment effect within each leaf. By separating these steps, GRF mitigates overfitting and reduces the bias in treatment effect estimates. It also uses adaptive weighting and subsampling to reduce bias and variance in treatment effect estimation. This technique ensures observations are weighted based on their similarity within the forest structure, ensuring that treatment effects are estimated locally and more precisely.

There is one final aspect of GRF which makes it a useful model for this thesis. GRF is flexible in handling complex, high-dimensional covariates and non-linear relationships. This adaptability makes GRF an effective tool for uncovering subtle patterns in the dataset which is used.

2.3.5 Bayesian Ridge

The final chosen model is a Bayesian Ridge Regression, a probabilistic extension of traditional ridge regression. Standard ridge regression uses an L2 penalty to address multicollinearity and overfitting. Bayesian Ridge Regression incorporates this method within a Bayesian framework, which allows uncertainty estimation in the model parameters (Tipping, 2001).

In this research, the regression coefficients are assumed to follow a prior Gaussian distribution centred at zero. The model updates these priors using observed data to obtain posterior distributions. This method provides measures of uncertainty for each coefficient, which can be beneficial when dealing with multicollinearity or limited data.

Bayesian Ridge Regression not only helps control overfitting by shrinking coefficients toward zero, similarly to classical ridge regression, but also provides probabilistic interpretations of the coefficients. This feature enhances the interpretability and allows for more informed decision making based on the results.

The implementation used in this research originates from the scikit-learn package, which is based on the algorithm described by Tipping, 2001.

Chapter 3

Method

3.1 Data preparation

In order to work with the supplied dataset, some data preparation was necessary. Given the nature of the dataset and certain limitations in it's structure, careful decisions were made regarding the handling of specific variables. In this section, the steps that have been taken will be explained.

3.1.1 Grouping data

Firstly, some variables contained inexplicable differences between the first and later waves of questioning. To avoid getting inconsistencies and to ensure uniformity across obervations, the decision has been made to only include the first recorded value.

Besides removing certain variables and values, some values have been grouped differently. To streamline the classification, the choice has been made to regroup these variables into larger groups. This will help reducing the sparsity, simplify the analysis, and make the categories more analytically meaningful while still retaining the information.

3.1.2 Missing data

To address missing values in the dataset, several imputation techniques were considered and applied. One of these methods is forward filling, where missing values are replaced by the most recent non-missing observation. Other variables containing missing data were assessed on a case-by-case basis. Finally, the choice has been made to drop the original variable for the irrigated land and exchange it for a binary format, only indicating the presence or absence of irrigation. This step was taken to combat the missing data in this variable, and to improve model compatibility and interpretation.

Multiple Imputation by Chained Equations (MICE) has been considered as an alternative strategy. However, this technique relies on strong assumptions about the data and the relationship between variables. When looking at the results,

we can see that the variables in this dataset are most likely not very correlated. Because of this, MICE would not have been the best imputation technique choice.

3.1.3 Deleting columns

Certain variables have been removed from the dataset due to their limited utility or high missing values count. One of these variables is a variable containing information on how the villagers were introduced to the insurance programme. This variable contained a large number of missing entries, leading to its deletion. Besides this, multiple variables related to phone ownership were substituted for one variable containing the information to avoid redundancy.

3.1.4 Encoding the data

Transforming the outcome variable

To address the issue of skewness in the outcome variable, a transformation was applied in order to improve the distribution and ensure better model performance. After testing multiple different transformation methods, the final choice landed on the **log(x+1)** transformation. The results of the log(x+1) transformation can be found in 3.1 and 3.2.

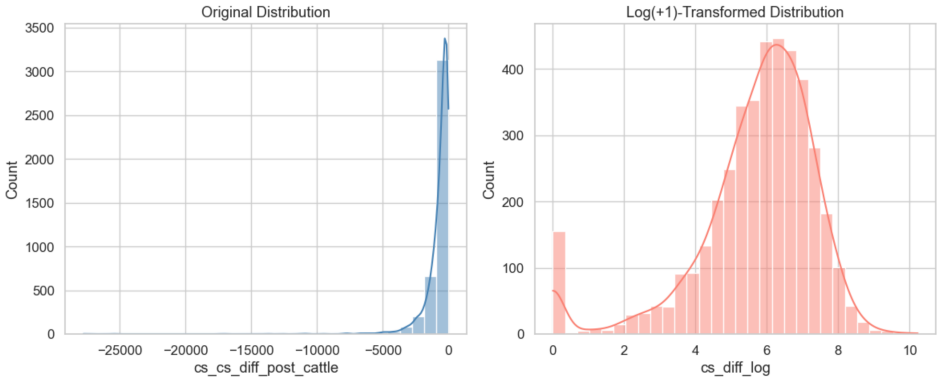

Figure 3.1: Distribution of the outcome variable before and after log(+1) transformation in the cattle dataset.

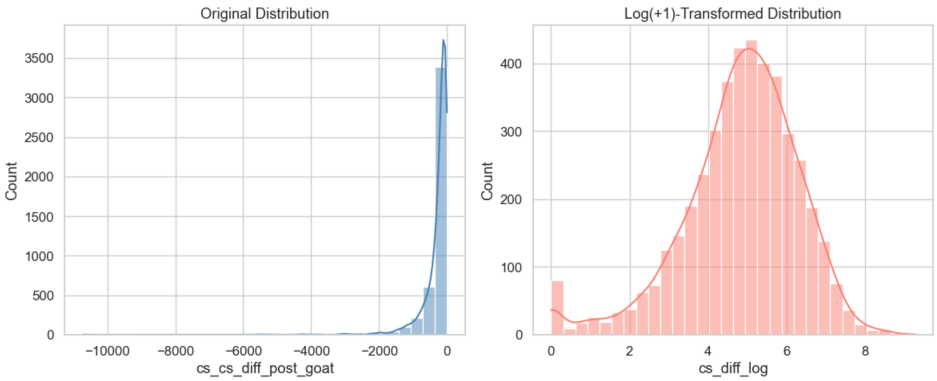

Figure 3.2: Distribution of the outcome variable before and after log(+1) transformation in the goat dataset.

Encoding the data

Before the models are ready to be trained, further preprocessing of the dataset is necessary. The final step consists of encoding the categorical variables in the dataset. This was done using one-hot encoding, which converted the categorical variables into binary indicator variables. This allows models to interpret categorical data without imposing ordinal relationships.

3.2 Models

In this research, several different models have been deployed in order to see which model performs best in predicting the impact of livestock shocks on household welfare. The selected models were chosen based on their ability to capture sparsity, nonlinear interactions, and heterogeneous treatment effects. Each of the models was tuned using a form of grid search or Bayesian optimisation, and their performance was evaluated using R^2, RMSE, and MAE on the test set.

1. R^2 measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. This metric ranges from negative infinity to 1.

2. **RMSE** is the square root of the average squared differences between predicted and actual values.

3. **MAE** calculates the average of the absolute differences between predicted and true values.

A short summary of the work on the different models is provided below.

3.2.1 Lasso Regression

In this implementation of Lasso, a standard Lasso and a form of grouped Lasso have been performed. For the singular Lasso implementation, grid search has been

performed to identify the optimal regularisation parameter α, balancing model simplicity and predictive accuracy.

For the group Lasso, the choice has been made to group the dummy variables of the categorical variables in the dataset. After this, a different grid search has been performed. For this version, a grid search was run on the *group_reg_value*, which is a parameter for the amount of penalty on different groups. The other tuned parameter is the *l1_reg_value*, which is a value that decides the penalty on individual features.

The model was fit on standardized data to ensure good comparability between the coefficients. To mitigate overfitting, cross-validation was used. This is also useful when evaluating the generalizability of the model across different data folds.

3.2.2 TabTransformer

For this research, categorical variables were embedded and passed through multiple transformer blocks with multi-head self-attention. This is a mechanism that enables the model to weigh the importance of each input feature relative to others, enabling the network to learn complex interactions between variables. Continuous variables have been standardized and concatenated before passing through a fully connected head for regression.

The hyperparameter tuning of this model has been done in two different ways: random search and Hyperband. Random search is an approach which randomly samples from a defined hyperparameter space and evaluates each configuration multiple times. This model has been trained using 15 combinations, 2 executions per trial, and 15 epochs per run.

The second strategy is called Hyperband. Hyperband adaptively allocates resources towards promising configurations while eliminating less effective ones. This method used a maximum of 50 trainings per epoch, and a reduction factor of three, meaning that each bracket trains fewer models for more epochs.

Both strategies explored a range of training parameters, including the number of transformer blocks, attention heads, hidden units, dropout rates, and learning rates.

3.2.3 Generalized Random Forest

To evaluate the predictive performance of the GRF model, the *grf* package in R was used. Using a wide set of explanatory variables, including continuous and categorical features, the model has tried to predict the target variable. The categorical values have been one-hot encoded using *model.matrix()*, which allowed the GRF to process them as numerical inputs.

Hyperparameter tuning was performed using a built-in function called *regression_forest()*, where all tunable parameters have been explored over 1000 repetitions and 1000 draws, with 100 trees per tuning round. The final model has been trained using 5000 trees to ensure stability in the predictions. After this, the model performance was assessed using out-of-bag predictions, with the same performance measures used with the other models.

3.2.4 Bayesian Ridge

Finally, an implementation of a Bayesian Ridge has been done. For this research, the *BayesianRidge* implementation from sklearn was used. To encode the categorical features, the *LabelEncoder* function in Python was applied. The continuous features were standardised.

After running this model with a basic setup of hyperparameters, grid search was applied to find the best-performing set of hyperparameters. Different combinations of the hyperparameters $\alpha 1$, $\alpha 2$, $\lambda 1$, and $\lambda 2$ were applied to explore different prior assumptions about the precision of weights and noise. $\alpha 1$ and $\alpha 2$ control the prior distribution over the noise precision (the inverse of the variance of the error term), $\lambda 1$ and $\lambda 2$ control the prior distribution over the precision of the weights (the size of the regression coefficients).

Chapter 4

Results

This chapter presents the empirical findings of the study, divided into three main sections. First, *Model Performance* compares the predictive accuracy of the four implemented models across the two datasets using $\boldsymbol{R^2}$, **MAE**, and **RMSE**. Next, *Hyperparameter Optimization* details the selected hyperparameters for the best-performing models, providing guidance in picking and tuning the most suitable model possible. Finally, the *Feature Importance* section examines the key predictors identified by each model, highlighting the most recurring variables across the datasets and models and discussing their relevance. Together these results offer an extensive evaluation of model performance and interpretability.

4.1 Model Performance

Model	**Dataset**	$\boldsymbol{R^2}$	**MAE**	**RMSE**
Lasso	Cattle	**0.011**	1.204	1.656
TabTransformers	Cattle	-0.006	1.286	1.783
GRF	Cattle	0.006	1.238	1.717
BayesianRidge	Cattle	-0.007	**1.203**	**1.655**
Lasso	Goat	0.021	1.124	1.499
TabTransformers	Goat	**0.046**	**1.056**	**1.412**
GRF	Goat	0.011	1.115	1.472
BayesianRidge	Goat	0.016	1.127	1.502

Table 4.1: This table shows the model performance comparison of all four models. The best performance metrics for both datasets have been bolted.

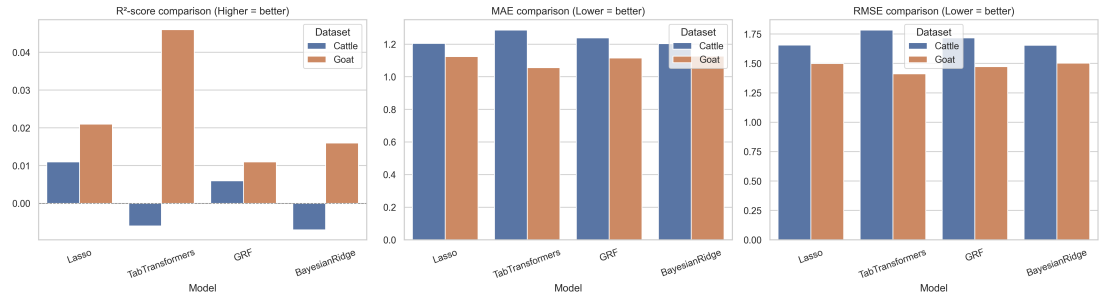

Figure 4.1: Visual model comparison overview. This overview clearly shows the difference in predictive accuracy between the two datasets.

Table 4.1 presents the performance of the four regression models applied on both the cattle and goat datasets. For each of the models, the R^2, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are reported.

Figure 4.1 shows the visual comparison between the models.

When basing the model performance on the R^2, the best-performing model on the cattle dataset is the Lasso model. The Lasso model outperformed all of the other models, where both the GRF and the TabTransformer models scored a higher R^2 score and lower MAE and RMSE. The Bayesian ridge model did manage to score the highest MAE and RMSE scores for this particular dataset.

When looking at the goat dataset, the TabTransformers model outperforms all of the other models. This model achieved the best scores across all models, even when compared to the other dataset. The difference between the performance of the TabTransformer model on the two datasets is big. All of the models did perform better on the goat dataset when compared on all three metrics. One possible explanation for the difference in performance is that goat-related welfare outcomes may follow more consistent patterns with less noise and stronger links to household characteristics. Additionally, the socio-economic variables in the dataset may be more predictive of goat-related insurance behaviour. These aspects could have let to the difference in model performance.

4.1.1 Hyperparameter Optimization of the best-performing models

Below, the hyperparameters of the best-performing models are noted.

Lasso

Hyperparameters for the best-performing model on the goat dataset: **l1_reg:** 1.00e-04

Hyperparameters for the best-performing model on the cattle dataset: **l1_reg:** 1.00e-02

TabTransfomers

Hyperparameters for the best-performing model on the goat dataset: **hidden_dim:** 128, **ff_dim:** 128, **dropout_rate:** 0, **num_heads:** 2, **learning_rate:** 0.001.

Hyperparameters for the best-performing model on the cattle dataset: **hidden_dim:** 32, **ff_dim:** 192, **dropout_rate:** 0, **num_heads:** 8, **learning_rate:** 0.0001.

GRF

Hyperparameters for the best-performing model on the goat dataset (rounded to three decimals): **sample.fraction:** 0.318, **mtry:** 18, **min.node.size:** 30, **honesty.fraction:** 0.754, **honesty.prune.leaves:** 0, **alpha:** 0.008, **imbalance.penalty:** 0.972.

Hyperparameters for the best-performing model on the cattle dataset (rounded to three decimals): **sample.fraction:** 0.444, **mtry:** 21, **min.node.size:** 53, **honesty.fraction:** 0.769, **honesty.prune.leaves:** 1, **alpha:** 0.003, **imbalance.penalty:** 0.439.

Bayesian ridge

Hyperparameters for the best-performing model on the goat dataset: $\alpha 1$**:** 1, $\alpha 2$**:** 1e-05, $\lambda 1$**:** 1e-05, $\lambda 2$**:** 1.

Hyperparameters for the best-performing model on the cattle dataset: $\alpha 1$**:** 0.01, $\alpha 2$**:** 1, $\lambda 1$**:** 1, $\lambda 2$**:** 0.0001.

4.2 Feature importance

In this section, the coefficients of the feature importance for the different models are discussed. The figures that accompany this information can be found in the appendix in chapter 5.3.

4.2.1 Lasso

Cattle

Figure B.1 shows the feature importances of the 20 most important features according to the Lasso model on the cattle dataset. The figure shows that the dummy variable corresponding to the household description **Not settled** is the most important feature. At second place, having a slightly lower coefficient, **Fully settled**, a dummy for the same variable, can be found. This shows the importance of this variable. Another variable that should be mentioned is the `trust_vip` variable. Both dummies, 'yes' and 'no', can be found in the top 10, showing the importance of this variable as well.

Goat

Figure B.2 shows the feature importances of the 20 most important features according to the Lasso model on the goat dataset. Similar to the cattle dataset, the variable concerning the household description with the dummy **Fully settled** scores very high. On this dataset it is even considered as the most important one. Besides this variable, the variable `trust_vip`, with dummy **No**, scores very high as well. Contrary to this dummy, the **Yes** dummy seems to get a score of 0.000.

4.2.2 TabTransformers

TabTransformers uses a slightly different method to show the feature importance. In this case, the permutation importance has been used. This score shows how much the R^2 decreases when permuting that feature.

Cattle

The most important feature, according to figure B.7, is a feature called `Numerical`. After further investigation, this feature is a clustering of the following features: age_constant, expend, irrigated_land_bin, number_minors, number_adults, and owns_phone. The importance of these features and their effect on the R^2 is 10 times as high as the second most important feature. The second most important feature is the feature called `educ_recoded_constant`, telling the user something about the level of education that has been achieved. One more variable worth mentioning once more is the `trust_vip` variable. Although it achieved a low score, it still showed to have some effect on the performance of the model.

Goat

Figure B.8 shows the same trend of the model on the goat dataset. `Numerical`, once more, is by far the most important feature. `trust_vip` is 5 times more important when compared to the importance in the cattle dataset. Another important feature is the `why_not_purchase_recoded`, scoring much higher on this dataset when compared to the cattle dataset. Finally, similar to the Lasso model, `household_description` shows to be an important feature when using the TabTransformers model, ranking fourth in importance.

4.2.3 GRF

Cattle

When looking at figure B.3, it is clear to see that the `Age_constant` is the most important feature in this scenario. Besides this feature, a few of the features we have seen before reappear. This model considers the `household_description`, and in this case, the **partially settled** dummy to be important features as well. Feature number four is a feature that is seen often aswell, namely `trust_vip`, and in this case the **yes** dummy. One final noticeable feature is `number_adults`, being

feature number three on the list and only having slightly less impact compared to the household description.

Goat

Figure B.4 shows the feature importances of this model on the goat dataset. Once again, the `household_description` variable is the most important variable. This time, the **'Not settled'** dummy is the most important factor. Besides this, another variable worth mentioning is `age_constant` which is not as important in this dataset but is still variable number three in the list. Showing up once more as well is the `trust_vip` variable. This variable seems to show up for all of the models and is clearly an important factor in this dataset.

4.2.4 Bayesian Ridge

Cattle

The results of the Bayesian ridge model on the cattle dataset can be found in Figure B.6. This figure shows a surprising feature at the top called `why_not_purchase`. Besides this feature, a few features that have been mentioned before show up. `household_description` and `trust_vip` are both in the top 5 most important features for this model. Another feature worth mentioning is the `irrigated_land_bin`, the self-made binary variable showing if the household has got irrigation systems.

Goat

Figure B.5 shows the feature importance on the goat dataset. This graph summarises what we have seen in the other models. `trust_vip` is the most important feature, being followed by `household_description`. It is clear that these two variables are important in all of the models. Besides these variables, this model gives high importance to the different language variables included in the model.

4.2.5 Summary of findings

Across all models and datasets, the `household_description` variable consistently emerges as a key predictor, highlighting the importance of settlement status. The variable `trust_vip` also shows a strong importance, indicating its influence on the outcomes. While the exact ranking varies between the different models and datasets, clear recurring themes can be found in these results. These results underline the multifaceted nature of the data and emphasise the value of combining different models and modelling approaches to capture these aspects of feature importance.

Chapter 5

Discussion

5.1 Conclusion

This thesis aimed to evaluate and compare the predictive performance and interpretability of four different regression models – Lasso, TabTransformers, Generalized Random Forest (GRF), and Bayesian Ridge Regression – on two datasets: the decision to purchase or not purchase IBLI for cattle and goats. The goal of this research was to understand which modelling approaches are best suited to make predictions on this dataset and to identify the key indicators contributing to these predictions. The results from this research help answer the two questions this thesis is based around:

1. **How effective are machine learning methods in predicting household welfare outcomes using observable socio-economic characteristics?**

2. **What are the most important household characteristics used to predict welfare on the IBLI dataset?**

5.1.1 Key findings

Model performance

The comparative analysis of model performance has provided several key insights. The first insight is the difference in model performance when comparing the two datasets. Although all models showed relatively low predictive power on the cattle data set, having a R^2 range from -0.007 to 0.011, performance was significantly better on the goat dataset, where the R^2 reached up to 0.046. This could indicate that the goat-related features could be more strongly associated with the target variable. It could also mean that the underlying patterns in this dataset were more easily learnable.

The Lasso regression turned out to be the top-performing model for the cattle dataset, achieving an R^2 score of 0.011 and having RMSE (1.656) and MAE (1.204) values that were slightly lower compared to the Bayesian Ridge model (RMSE = 1.655, MAE = 1.203). The Bayesian Ridge model had a slightly lower R^2 score

but it can be concluded that both linear models were comparable in performance on this particular dataset.

TabTransformers outperformed all of the other models on the goat dataset. The model achieved the best scores for all three metrics: R^2 (0.046), MAE (1.056), and RMSE (1.412). This result suggests that for more structured datasets with higher signal-to-noise ratios, deep learning models can capture non-linear interactions more effectively than other model types.

Feature importance

A close inspection of the results of the importance of the characteristics has revealed strong commonalities between models and datasets. First of all, the household settlement status consistently emerged as a top predictor. Although the dummy variables varied between `Not settled`, `Partially settled`, and `Fully settled`, it appears that this variable has proven very insightful to the models. It also shows the importance of the household settlement status and the welfare variable.

Besides the household settlement status variable, `trust_vip` also appeared in the top features across all models, reaffirming its central role in determining the welfare outcome variable. This variable displays the trust people have in the village insurance promoter. The fact that this variable also scores very high across all models could reflect the relevance of social trust or institutional confidence in shaping economic decisions.

TabTransformers managed to highlight the significance of aggregated indicators, particularly the engineered feature `Numerical`. This feature encompassed multiple variables like `age`, `land ownership`, `household composition`, `and phone ownership`. This result suggests that composite features may provide a broader image to machine learning models.

Models like the GRF and Bayesian Ridge seemed to profit more from the more traditional socio-demographic variables like `age_contstant`, `number_adults`, and `educ_constant`, reflecting patterns that are consistent with already established domain insights.

The first research question can be answered in a slightly disappointing way. The trialled machine learning models have shown very modest predictive power, although they are showing better performance on the goat dataset. Although linear models like Lasso and Bayesian Ridge performed reasonably well on the likely weaker cattle data, the TabTransformers excelled on the goat dataset. This suggests that deep learning models are more effective when the data is richer and contains non-linear patterns. In general, machine learning models can be useful in predicting welfare outcomes, but their effectiveness and performance depend heavily on data quality and data size.

The most important features that are used to predict on the IBLI dataset include the `household_settlement_status`, and the `trust_vip`, both consistently ranking high across models and datasets. TabTransformers highlighted the importance of composite features, while GRF and Bayesian Ridge leaned more on

the traditional socio-demographic indicators (age, phone ownership, number of adults in a family). These findings confirm the importance of both individual socio-economic variables and aggregated indicators in predicting welfare.

5.1.2 Implications

By leveraging the longitudinal data which has been gathered, this study demonstrates that even modest predictive models provide valuable insight into the household welfare dynamics. The models have been trained on a comprehensive dataset reflecting herd dynamics, household behaviours, and more, making the findings of this research relevant to pastoralist welfare and the prediction of this welfare.

The moderate importance of features like settlement status and social trust suggest that these observable household characteristics carry meaningful signals. These insights could help stakeholders design more targeted interventions, like outreach campaigns or subsidy allocations, tailored to households that are predicted to benefit most from IBLI. In practical terms, the ability to identify the important features could enable policy makers to refine IBLI schemes. This would not only improve coverage but also increases the program's overall efficiency and fairness.

Besides these practical, real-life contributions, this thesis also bridges machine learning with traditional demographic modelling in a context that has real-world implications. Based on these results, a recommendation on the usage of machine learning models can be made. If the dataset will remain the same, it is useful to make use of the simpler models like the Lasso model, since it will perform better on small datasets. TabTransformers does show promising results when dealing with data that contains relationships between the variables that could be harder to find. When improving or expanding this dataset, TabTransformers could prove to be a useful model in predicting feature importance and welfare outcomes.

In conclusion, this research contributes not only to the methodological literature available but also to the applied field of rural development. Using a data set grounded in years of fieldwork, it provides relevant insights into the understanding of supporting sustainable livelihoods.

5.2 Limitations

This research has had to deal with certain limitations. The main source of limitations was the dataset. The size of the dataset was not ideal for training machine learning models, especially for the data-intensive neural network models like TabTransformers. The effects can be seen in the model performances of the different models. Although some feature importance information could be collected, the performance of the models is very modest. Besides the size of the dataset, the dataset also needed some heavy preprocessing since it experienced a lot of issues with missing data. After this preprocessing, a dataset which could be properly worked with was made, but this process does come at an expense. It can influence the outcomes, making them biased or less generalisable.

The next issue was the absence of computational power. It is possible that larger, more powerful models would have performed better on the dataset, but this type of power was not available. Since these experiments and models have been run locally, the models were limited in size.

5.3 Future research

The final limitation of this research opens up interesting opportunities for future research. With the availability of more computational power, larger, smarter, and newer models can be tried on this dataset. This thesis can be a step in the right direction, especially when looking at the feature importances. This information can be used in new research using different models.

Besides adding to the amount of computational power, adding information to the datasets could open up new research possibilities. Filling in missing information, adding rounds of surveying, and adding households to the dataset can help models improve their predictive power. If enough data over more time can be collected, treatment-effect research can be done on the welfare of the households who have purchased IBLI compared to the households who have not.

Bibliography

Amazon Web Services. (2024). Tabtransformer — how it works [Accessed: 2025-06-11].

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Barrett, C. B., Harrison, G. W., Jensen, N., Morsink, K., Schneider, M., Swarthout, J. T., & Upton, J. (2019). Do no harm: Evaluating the welfare effects of behavioral insurance interventions in ethiopia. *Journal of Economic Behavior Organization*, *162*, 167–185. https://doi.org/10.1016/j.jebo.2019.03.017

Chantarat, S., Mude, A. G., Barrett, C. B., & Carter, M. R. (2008). *Index-based livestock insurance for northern kenya's arid and semi-arid lands: The marsabit pilot* (tech. rep.). ILRI.

Chantarat, S., Mude, A. G., Barrett, C. B., & Turvey, C. G. (2017). Welfare impacts of index insurance in the presence of a poverty trap. *World Development*, *94*, 119–138. https://doi.org/10.1016/j.worlddev.2016.12.044

Harrison, G. W., Morsink, K., & Schneider, M. (2020). *Do no harm? the welfare consequences of behavioural interventions* (tech. rep. No. 2020-12) (Available at: https://cear.gsu.edu/files/2020/08/WP_2020_12_Do_No_Harm.pdf). Center for Economic Analysis of Risk (CEAR) Working Paper.

Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*. https://arxiv.org/abs/2012.06678

International Livestock Research Institute (ILRI). (2025). About us – index-based livestock insurance (ibli) [Accessed June 10, 2025].

Jensen, N., Barrett, C. B., & Mude, A. G. (2017). Designing index-based livestock insurance for managing asset risk in northern kenya. *Journal of Risk and Insurance*, *84*(1), 123–149.

Jensen, N. D., Barrett, C. B., & Mude, A. G. (2015). *The favourable impacts of index-based livestock insurance: Evaluation results from ethiopia and kenya* (ILRI Research Brief No. 52). International Livestock Research Institute (ILRI). Nairobi, Kenya. https://hdl.handle.net/10568/66652

Jensen, N. D., Fava, F. P., Mude, A. G., Barrett, C. B., Wandera-Gache, B., Vrieling, A., Taye, M., Takahashi, K., Lung, F., & Ikegami, M. (2024). *Escaping poverty traps and unlocking prosperity in the face of climate risk: Lessons from index-based livestock insurance*. Cambridge University Press.

Karlan, D., Osei, R., Osei-Akoto, I., & Udry, C. (2014). Agricultural decisions after relaxing credit and risk constraints. *The Quarterly Journal of Economics*, *129*(2), 597–652. https://doi.org/10.1093/qje/qju002

Takahashi, K., Ikegami, M., Sheahan, M., & Barrett, C. B. (2016). Experimental evidence on the drivers of index-based livestock insurance demand in southern ethiopia. *World Development*, *78*, 324–340. https://doi.org/10.1016/j.worlddev.2015.10.039

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, *1*(Jun), 211–244.

Appendix A

Variable explanation

Table A.1: Variable Overview and Descriptions

Variable Name	**Description**
afm_language	Binary variable showing whether the participant speaks the Afan Omoro language
age_constant	Age of the participant
agric_land	Binary variable showing if the participant has agricultural land
amh_language	Binary variable showing if the participant speaks Amharic
educ_recoded_constant	Variable showing the education level of the participant
eng_language	Binary variable showing if the participant speaks the English language
irrigated_land_bin	Binary variable showing if the participant has irrigated land
cs_cs_diff_post	Outcome variable showing computed welfare measure
number_minors	The number of minors in a household
educ_child_recoded	Variable showing the level of education of the child
activity_child_recoded	Variable showing the main activities of the child, like work or school
household_description	Variable showing whether a household is settled, partially settled, or not settled at all
number_adults	The number of adults in a household
main_info_source_recoded	Variable showing the main information source of the participant about IBLI
religion_recoded	Variable showing the religion the participant belongs to
owns_phone	Binary variable showing if the participant owns a phone

Variable Name	Description
household_moved	Binary variable showing if a household has moved in the last 6 months
why_not_purchase_recoded	Variable showing the reason why a participant did not purchase IBLI
know_vip	Variable showing if the participant knows the Village Insurance Promoter
trust_vip	Variable showing if the participant trusts the Village Insurance Promoter

Appendix B

Feature Importance

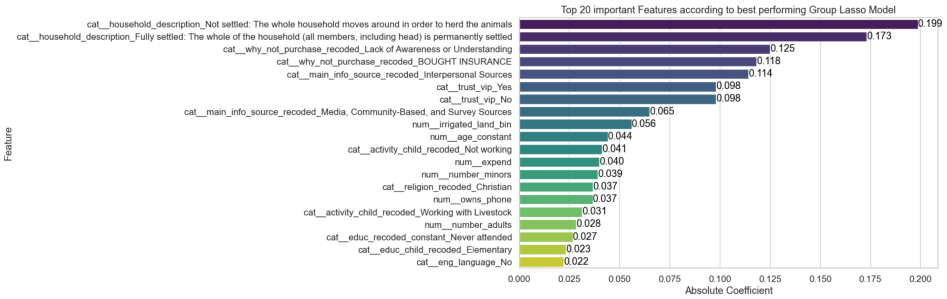

Figure B.1: Feature importance according to the best-performing Lasso model on the cattle dataset.

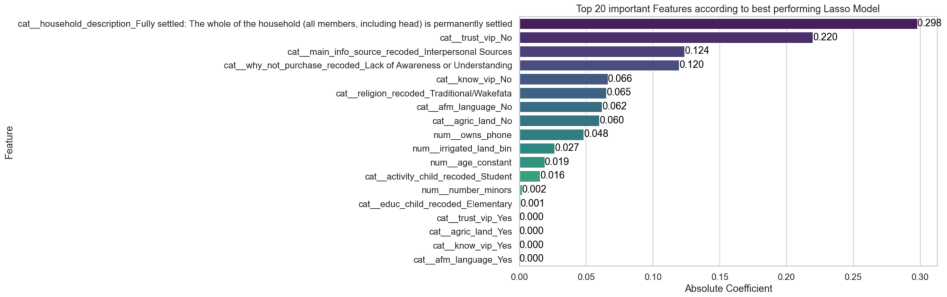

Figure B.2: Feature importance according to the best-performing Lasso model on the goat dataset.

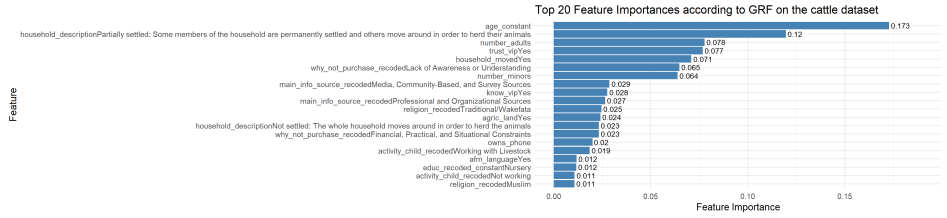

Figure B.3: Feature importance according to the best-performing GRF model on the cattle dataset.

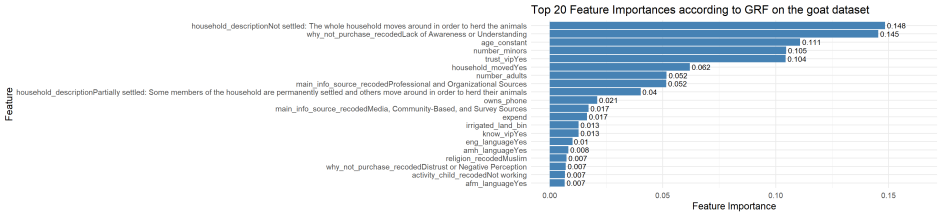

Figure B.4: Feature importance according to the best-performing GRF model on the goat dataset.

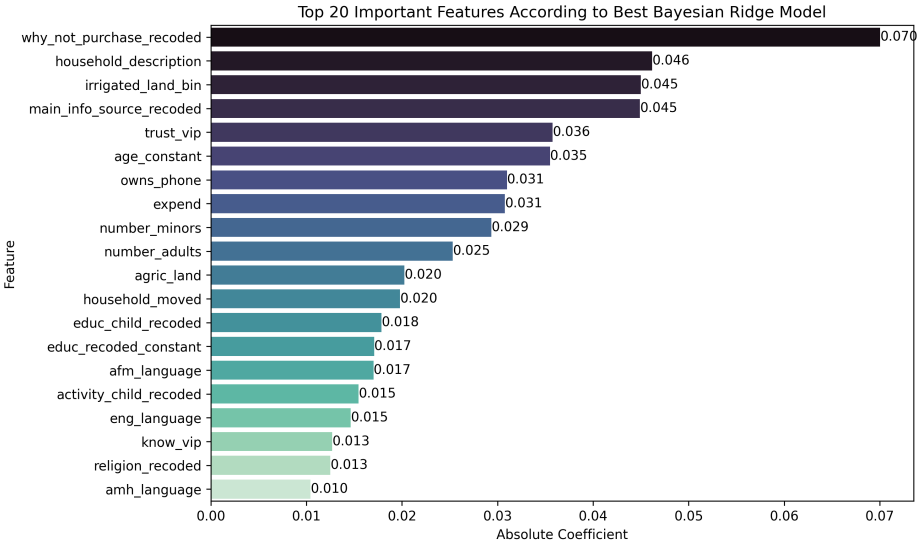

Figure B.5: Feature importance according to the best-performing bayesian ridge model on the goat dataset.

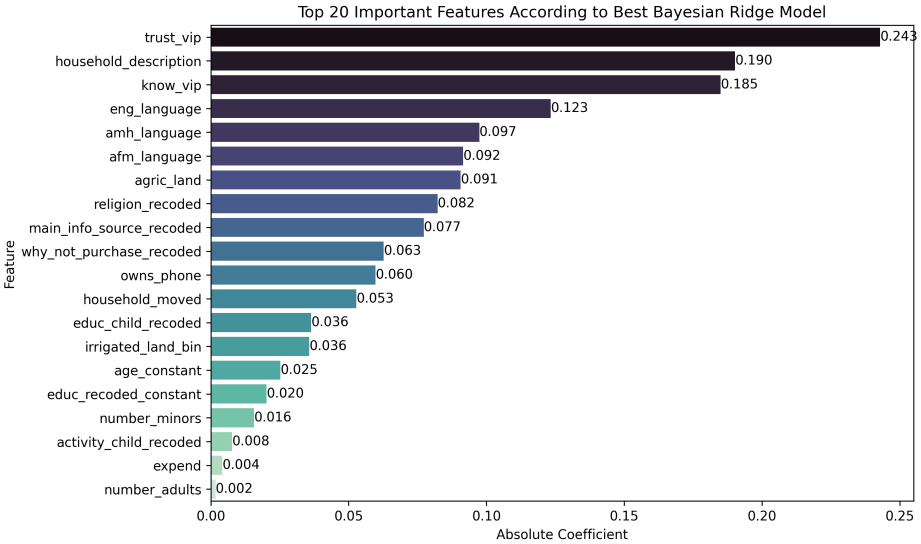

Figure B.6: Feature importance according to the best-performing bayesian model on the cattle dataset.

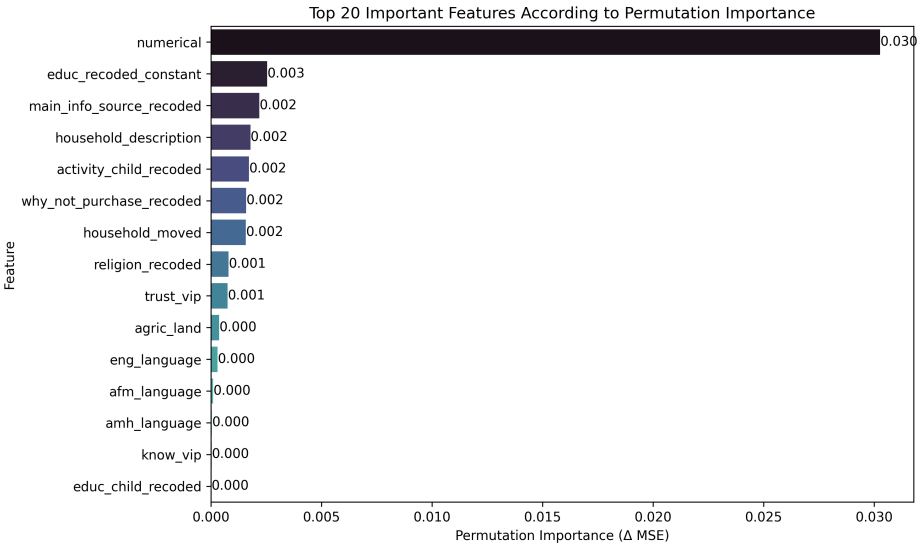

Figure B.7: Feature importance according to the best-performing TabTransformers model on the cattle dataset.

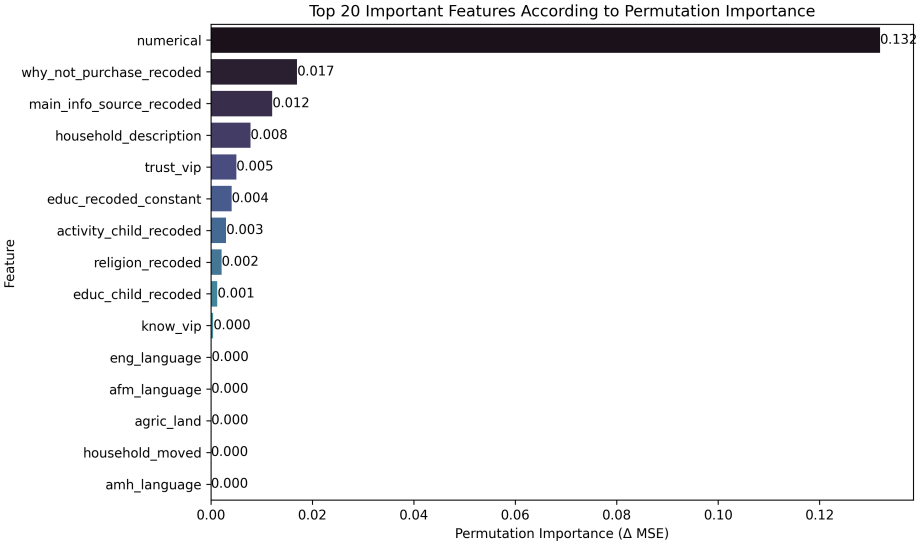

Figure B.8: Feature importance according to the best-performing TabTransformers model on the goat dataset.