Ordinal Multi-class Classification Problem on Harvard GLOPOP Data

Country of choice: Tajikistan

Author: Shahnoza Nurubloeva

Date: 08/07/2025

Data Acquisition

The GLOPOP-S dataset is a comprehensive global synthetic population database hosted on the Harvard Dataverse. The data was obtained from global synthetic population database GLOPOP-S. They took the Demographic and Health Surveys for 45 countries including Tajikistan, categorized and harmonized it, as well as went through other data processing steps including creating regional marginal distributions.

In case of missing data in the regions, to impute missing regional data, new values were modeled based on *k* number of regions with similar subnational human develepment index (SHDI) or SHDI relative to national HDI.

Then, marginals were scaled to match the region's population size in 2015 to conserve the shape of the marginal distributions.

The following methods were used to generate a synthetic population from DHS and LIS: synthetic reconstruction, combinatorial optimization, statistical learning.

All the columns of the data, although represented as numbers, are of categorical type. Below is the description of the columns taken from the article "A global dataset of 7 billion individuals with socio-economic characteristics" Nature journal.

Attributes	HH/I	Levels
Income	Individual	1: poorest 20%, 2: poorer 20%, 3: middle 20%, 4: richer 20%, 5: richest 20%, −1: unavailable for country
Wealth	Individual	1: poorest 20%, 2: poorer 20%, 3: middle 20%, 4: richer 20%, 5: richest 20%, −1: unavailable for country
Settlement type	Household	0: urban, 1: rural
Age	Individual	1: 0–4, 2: 5–14, 3: 15–24, 4: 25–34, 5: 35–44, 6: 45–54, 7: 55–64, 8: 65 +
Gender	Individual	1: male, 0: female
Education	Individual	1: less than primary, 2: complete primary, 3: incomplete secondary, 4: complete secondary or tertiary, 5: higher
Household type	Household	1: single, 2: couple, 3: couple with children, 4: one parent with children, 5: couple with (non-) relatives, 6: couple with children and (non)-relatives, 7: one parent with children and (non-) relatives, 8: other
Household ID	Household	1, …,
Relationship to head	Individual	1: head, 2: partner, 3: child, 4: relative, 5: non-relative
Household size	Household	1: 1, 2: 2, 3: 3–4, 4: 5–6, 5: 7–10, 6: 10+
Ownership of agricultural land (DHS)	Household	1: yes, 2: no, −1: unavailable for country
Floor material (DHS)	Household	1: natural, 2: rudimentary, 3: finished, −1: unavailable for country
Wall material (DHS)	Household	1: natural, 2: rudimentary, 3: finished, −1: unavailable for country
Roof material (DHS)	Household	1: natural, 2: rudimentary, 3: finished, −1, unavailable for country
Source	Country	1: LIS, 2: LIS survey, 3: LIS marginals, 4: Modeled by LIS data, 5: DHS, 6: Modeled by DHS data

Table 2. Attributes in GLOPOP-S.

Exploratory Data Analysis

General Observations

- The dataset contains **9,269,799 records** with **16 variables**, all of which are categorical (encoded as integers).
- **No missing values** are present in any column (`Miss = 0` for all variables).

Key Variable Insights

1. Household ID (`HID`)

- High duplication: ~20 records per household on average, suggesting longitudinal or multi-member sampling.

2. Household Relationships (`RELATE_HEAD`)

- Majority are neither head nor primary dependents (median=3), hinting at extended family structures.

3. Economic Indicators (`INCOME` , `WEALTH`)

- `INCOME` : The column is constant - entirely masked (all `1`) - required removal.
- `WEALTH` : Slight right skew (Q75=4), but few reach the highest tier (Q95+=5).

4. Geographic/Rural Status (`RURAL`)

- Only 19.5% rural.

5. Demographics (`AGE` , `GENDER` , `EDUC`)

- `AGE` : 8 age groups. Mean ~3.6 (median = 3), suggesting a younger population.
- `GENDER` : Near-balanced (mean = 0.496, ~49.6% female).
- `EDUC` : 5-tiered education levels. Mean ~2.96 (median = 3), indicating most have mid-level education.

6. Household Structure (`HHTYPE` , `HHSIZE_CAT`)

- `HHTYPE` : Has complex non-ordinal categories. No conclusions based on numeric representation of groups.
- `HHSIZE_CAT` : Median = 5 out of 6 categories, mean = 4.54, suggesting larger households.

7. Housing Quality (`FLOOR` , `WALL` , `ROOF`)

- `FLOOR` : Ordinal category of 3 types. Mean = 2.28 favoring "finished".
- `WALL` : Mean = 2.08, suggesting mixed types.
- `ROOF` : Strong skew toward category 3 (median = 3, mean = 2.96), indicating majority has "finished".

8. Agriculture (`AGR_CONNERSHIP`)

- **Binary (0/1)**: 52.1% have agricultural connections (mean = 0.521), reflecting Tajikistan's agrarian economy.

9. Region (`REGION`)

- **5 regions**: Mean = 2.86 (median = 3), with values spread across categories.

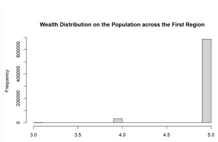
Exploratory Data Analysis (EDA)																			
	Type	Recs	Miss	Unique	Min	Q25	Q50	Avg	Q75	Max	StDv	Neg	Zero	Pos	OutLo	OutHi	Q95	Q96	Q97
HID	Num/Int	9 269 799	0	448 902	1	100686	206409	208566.117	313508	448902	125379.739	0	0	9 269 799	0	0	406766	411043	415501
RELATE_HEAD	Num/Int	9 269 799	0	5	1	2	3	2.807	4	5	0.994	0	0	9 269 799	0	0	4	4	4
INCOME	Num/Int	9 269 799	0	1	-1	-1	-1	-1.000	-1	-1	0.000	9 269 799	0	0	0	0	-1	-1	-1
WEALTH	Num/Int	9 269 799	0	5	1	2	3	2.979	4	5	1.407	0	0	9 269 799	0	0	5	5	5
RURAL	Num/Int	9 269 799	0	2	0	0	0	0.195	0	1	0.396	0	7 464 316	1 805 483	0	1 805 483	1	1	1
AGE	Num/Int	9 269 799	0	8	1	2	3	3.616	5	8	1.990	0	0	9 269 799	0	0	7	8	8
GENDER	Num/Int	9 269 799	0	2	0	0	0	0.496	1	1	0.500	0	4 673 581	4 596 218	0	0	1	1	1
EDUC	Num/Int	9 269 799	0	5	1	1	3	2.959	4	5	1.465	0	0	9 269 799	0	0	5	5	5
HHTYPE	Num/Int	9 269 799	0	8	1	3	6	5.088	6	8	1.598	0	0	9 269 799	0	0	7	7	7
HHSIZE_CAT	Num/Int	9 269 799	0	6	1	4	5	4.536	5	6	0.999	0	0	9 269 799	214 285	0	6	6	6
AGRI_OWNERSHIP	Num/Int	9 269 799	0	2	0	0	1	0.521	1	1	0.500	0	4 441 291	4 828 508	0	0	1	1	1
FLOOR	Num/Int	9 269 799	0	3	1	2	3	2.278	3	3	0.818	0	0	9 269 799	0	0	3	3	3
WALL	Num/Int	9 269 799	0	3	1	1	3	2.077	3	3	0.964	0	0	9 269 799	0	0	3	3	3
ROOF	Num/Int	9 269 799	0	3	1	3	3	2.953	3	3	0.301	0	0	9 269 799	227 294	0	3	3	3
SOURCE	Num/Int	9 269 799	0	1	5	5	5	5.000	5	5	0.000	0	0	9 269 799	0	0	5	5	5
REGION	Num/Int	9 269 799	0	5	1	2	3	2.863	4	5	1.012	0	0	9 269 799	0	0	4	4	4

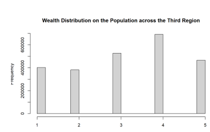
Data Split

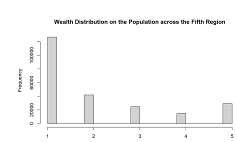
To develop the machine learning model, the dataset—comprising over 9 million observations—was randomly shuffled to eliminate any potential ordering bias. Following this, the data was partitioned into training and testing subsets using a 70:30 ratio based on index-based slicing. Given the large volume of data, stratified sampling was deemed unnecessary, as random shuffling was sufficient to ensure representative distribution across both subsets.

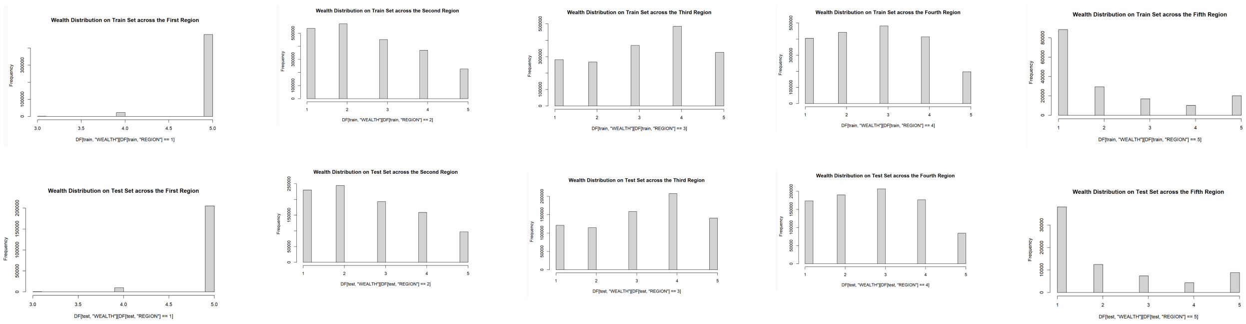
Homogeneity Across Regions and Data Splits Test

A homogeneity check was conducted to verify the consistency of data distribution across different regions and between the training and test subsets. This was done by constructing and analyzing histograms for key variables within each region and dataset split. The goal was to ensure that the random shuffling and partitioning process preserved the underlying statistical properties and regional characteristics of the data. The first row shows distribution of WEALTH variable across all five regions in the Population Set, while the second and the third rows — Train and Test Sets accordingly.

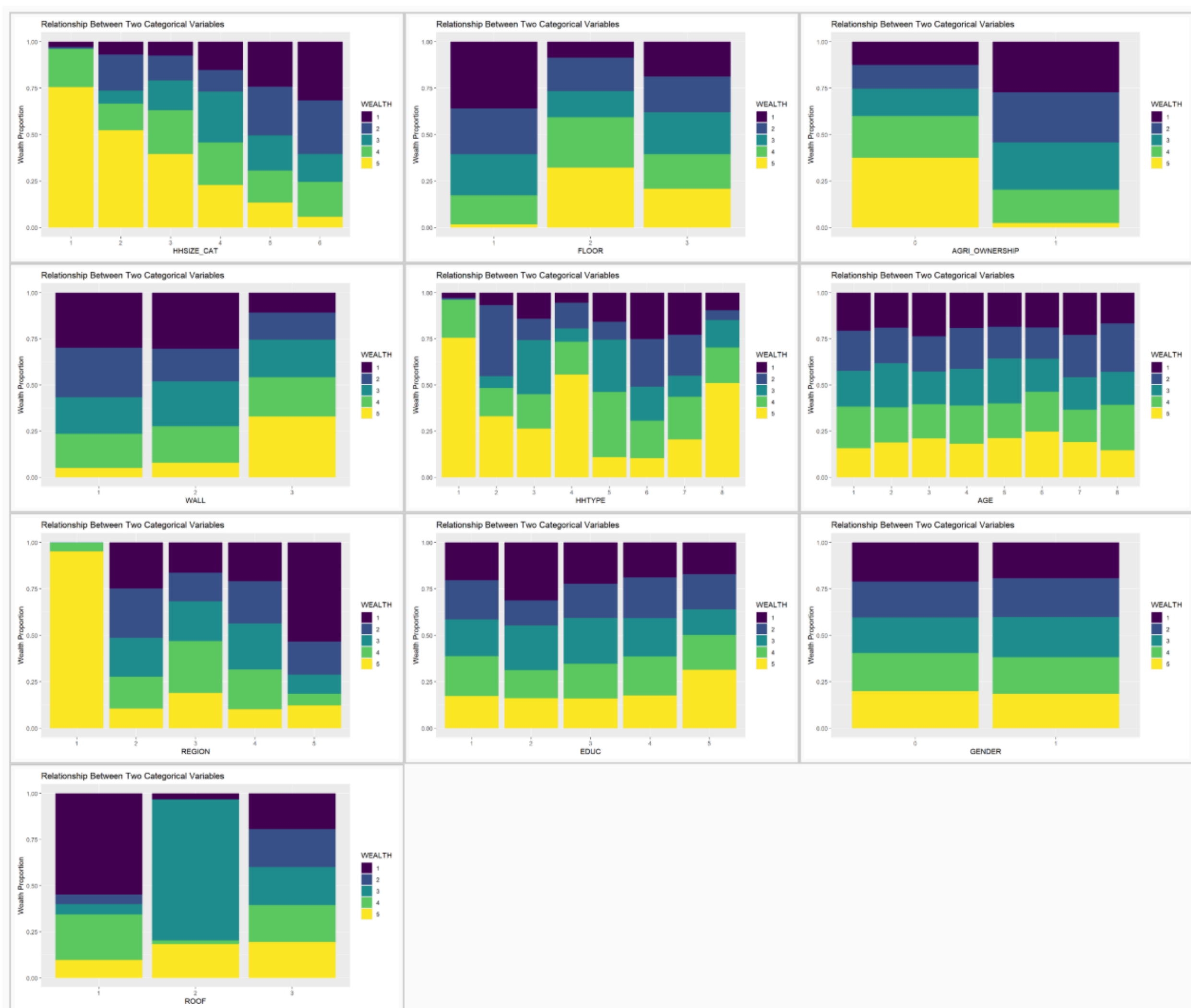








Relationships between WEALTH and other variables



Ordinal Logistic Regression Model

Model: MASS::polr()

Model Assumptions

1. **Proportional odds**: The relationship between each predictor and the outcome is consistent across all outcome thresholds.
2. **No multicollinearity**: Check VIF (<5).
3. **Adequate sample size**: ~10-20 cases per predictor per outcome level.
4. **Meaningful ordinal outcome**: Categories must follow a logical order.

Correlation Matrix

AGE	GENDER	EDUC	HHTYPE	HHSIZE_CAT	AGRI_OWNERSHIP	FLOOR	WALL	ROOF	REGION
1.00	−0.02	0.72	−0.04	−0.15	−0.02	−0.01	0.02	−0.07	0.02
−0.02	1.00	0.11	−0.07	−0.02	0.04	0.03	0.00	−0.02	0.00
0.72	0.11	1.00	−0.10	−0.20	−0.01	0.01	0.13	−0.08	−0.03
−0.04	−0.07	−0.10	1.00	0.50	0.04	−0.03	−0.01	0.15	0.02
−0.15	−0.02	−0.20	0.50	1.00	−0.04	−0.12	−0.18	0.36	0.00
−0.02	0.04	−0.01	0.04	−0.04	1.00	0.04	−0.14	−0.32	0.22
−0.01	0.03	0.01	−0.03	−0.12	0.04	1.00	0.06	−0.01	0.09
0.02	0.00	0.13	−0.01	−0.18	−0.14	0.06	1.00	0.14	−0.32
−0.07	−0.02	−0.08	0.15	0.36	−0.32	−0.01	0.14	1.00	−0.31
0.02	0.00	−0.03	0.02	0.00	0.22	0.09	−0.32	−0.31	1.00

Mutual Information

```
##                attr_importance
## AGE               0.0066949491
## GENDER            0.0008184556
## EDUC              0.0125136209
## HHTYPE            0.0659387786
## HHSIZE_CAT        0.0727038513
## AGRI_OWNERSHIP    0.1289645686
## FLOOR             0.0719929378
## WALL              0.0874742434
## ROOF              0.0108648835
## REGION            0.1495840274
```

Conclusion: Based on the Mutual Information table and Correlation matrix the following features can be removed: `AGE` and `HHTYPE`.

Model

A classical statistical model for ordinal outcomes that assumes proportional odds - the effect of predictors is consistent across all outcome thresholds. It provides interpretable coefficients but requires strict assumptions about data structure and linearity.

```
## Call:
## MASS::polr(formula = WEALTH ~ EDUC + HHSIZE_CAT + WALL + ROOF +
##     FLOOR + AGRI_OWNERSHIP, data = DF[1:1000000, c(features,
##     "WEALTH")])
##
## Coefficients:
##          EDUC.L          EDUC.Q          EDUC.C          EDUC^4    HHSIZE_CAT.L
##       0.1077399       0.2283148      -0.3086780       0.2799945      -2.2316220
##    HHSIZE_CAT.Q    HHSIZE_CAT.C    HHSIZE_CAT^4    HHSIZE_CAT^5           WALL2
##      -0.2713608      -0.5475459       0.4753573      -0.1532835       0.3280669
##           WALL3           ROOF2           ROOF3          FLOOR2          FLOOR3
##       1.3573465       0.4965558       1.2079452       1.7566316       1.2148722
## AGRI_OWNERSHIP1
##      -2.0293344
##
## Intercepts:
##        1|2        2|3        3|4        4|5
## -0.4812981  0.8579052  2.0798659  3.6655017
##
## Residual Deviance: 2626380.86
## AIC: 2626420.86
```

The Proportionality of Odds Assumption Test

```
## --------------------------------------------------
## Test for     X2  df  probability
## --------------------------------------------------
## Omnibus          100129.44   48  0
## EDUC.L           2034.8  3   0
## EDUC.Q           3126.64 3   0
## EDUC.C           2375.34 3   0
## EDUC^4           139.39  3   0
## HHSIZE_CAT.L     10546.93    3   0
## HHSIZE_CAT.Q     32189.8 3   0
## HHSIZE_CAT.C     1802.41 3   0
## HHSIZE_CAT^4     2792.69 3   0
## HHSIZE_CAT^5     957.51  3   0
## WALL2            883.39  3   0
## WALL3            20574.94    3   0
## ROOF2            7103.63 3   0
## ROOF3            31038.62    3   0
## FLOOR2           11596.78    3   0
## FLOOR3           12959.61    3   0
## AGRI_OWNERSHIP1  51492.14    3   0
## --------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

ull hypothesis regarding the proportional odds assumption is rejected for all predictors. Hence, non of the predictors are acceptable for the modeling.

Multicollinearity Test

```
##         Variable     GVIF Df GVIF^(1/(2*Df))
## 1     HHSIZE_CAT 1.223222  5        1.020353
## 2 AGRI_OWNERSHIP 1.134102  1        1.064942
## 3           WALL 1.109140  2        1.026234
## 4          FLOOR 1.100145  2        1.024148
## 5           ROOF 1.099745  2        1.024054
## 6           EDUC 1.086245  4        1.010395
```

Hence, no variable is significantly multicollinear.

Model Evaluation

→ Train Set

Metric	Value
Multi-class area under the curve	0.7417
Gini	0.483420297892295

→ Test Set

```
# Confusion Matrix
        Actual
Predicted     1     2     3     4     5
        0     0     0     0     0     0
        1 46554 46585 46966 45594 44588
        2 32089 32299 32909 31801 30709
        3 30592 29850 30082 29661 28696
        4 53309 53203 53395 52380 51139
```

Multi-class area under the curve	0.5002
Gini on test set	0.000441979905060563
Cohen's Kappa	-0.000370885631921676
Weighted Kappa	0.0007372
Ordinal Concordance	0.00036373355246857
Accuracy	0.199767946450715
Precision	0.287673601462297
Recall	0.202291054206273
F1 Score	0.237542832667047

📎 Conclusion on the model: Since the proportionality of odds assumption does not hold for this model, even though no other assumptions are violated, this model is in appropriate to be used.

Generalized Linear (Logistic) Regression Model

VGAMS::vglm()

A flexible framework for fitting various regression models, including extended ordinal logistic models that can relax the proportional odds assumption. This allows more flexibility and higher quality in compare to `polr`. Offers sophisticated modeling options like partial proportional odds and continuation ratio models.

As the Thomas W. Yee suggests in his article [The VGAM Package for Categorical Data Analysis](file:///C:/Users/SNurubloeva/Downloads/v32i10.pdf), "the framework is very well suited to many 'classical' regression models for categorical
responses".

Model Assumptions

1. **Proportional odds**: The assumption of Ordinal Logistic Regression is relaxed on a feature if it's included in family = cumulative(parallel = FALSE ~ .) of VGLM model.
2. **No multicollinearity**: Check VIF (<5).
3. **Adequate sample size**: ~10-20 cases per predictor per outcome level.
4. **Meaningful ordinal outcome**: Categories must follow a logical order.

Modification of predictors' bins due to non-homogenneity

```
DF$HHSIZE_CAT <- fct_collapse(DF$HHSIZE_CAT,
                              "1_2" = c("1", "2"),
                              "3_4" = c("3", "4"),
                              "5_6" = c("5", "6")) %>% factor(., ordered = TRUE)

DF$ROOF <- fct_collapse(DF$ROOF,
                        "1_2" = c("1", "2"),
                        "3" = c("3")) %>% factor(., ordered = TRUE)
```

```
model <- vglm(
  WEALTH ~ EDUC + AGRI_OWNERSHIP + FLOOR + WALL*ROOF + REGION + HHSIZE_CAT+ EDUC*AGRI_OW
  family = cumulative(parallel = FALSE ~ AGRI_OWNERSHIP + FLOOR + WALL),  # TRUE ~ EDUC + ROOF + RI
```

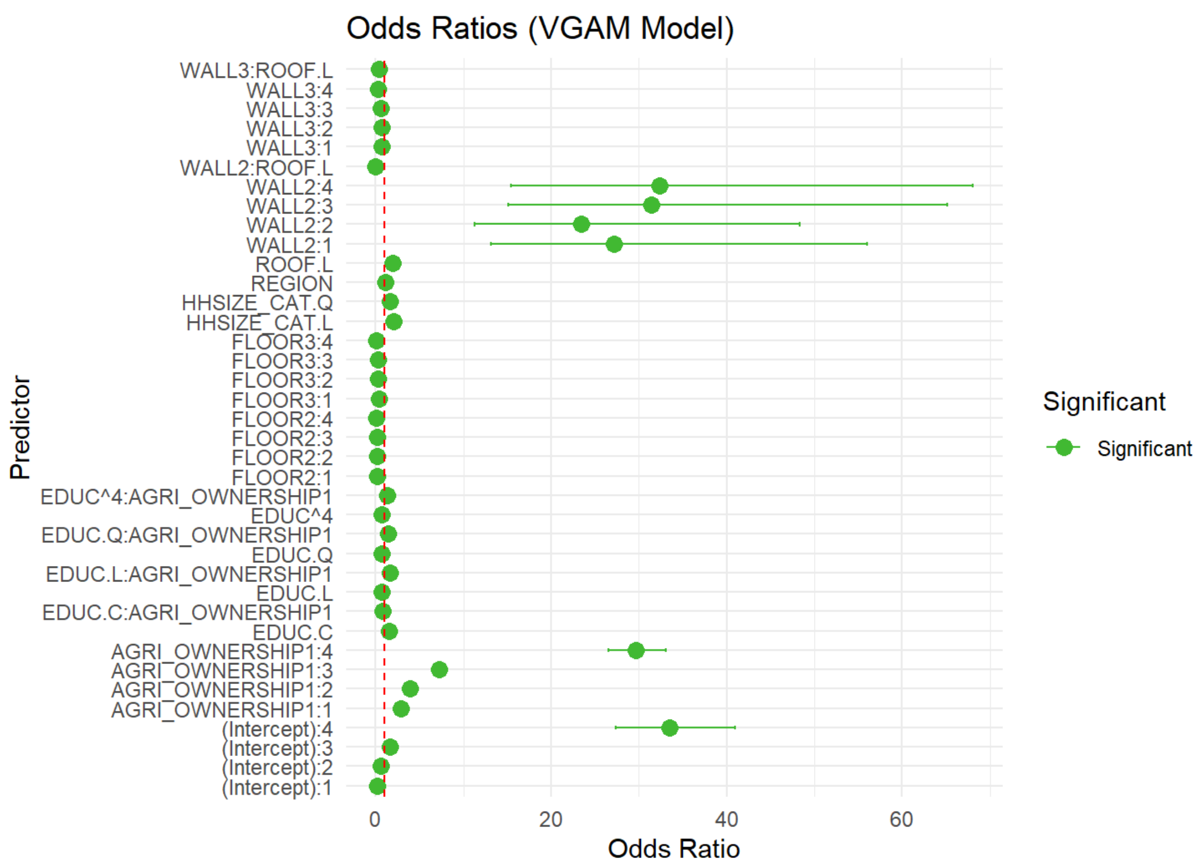
```
  data = DF[1:1000000, ]
)
```

> Relaxation of proportionality of odds assumption was performed not to all predictors but only to half of them.

Possible reasons:

- The model becomes overly complex when too many predictors are allowed to have non-proportional effects (each relaxed predictor gets **K-1 coefficients**, where **K = number of outcome categories**).
- If the data lacks sufficient variation or has small cell counts, the model may fail to converge.

Odds Ratio Test



Model Evaluation

→ Train Set

Multi-class area under the curve	0.759

Gini	0.518007

→ Test Set

```
# Confuction Matrix
         Actual
Predicted      1      2      3      4      5
        1 278925 156001  87674  59444   1980
        2  65302 113456  65225  33085  10692
        3 121724 106729 156987 107103  19277
        4  85771 166849 214373 255119  71785
        5  10087  17973  41309 102781 431290
```

Model Performance Metrics		
Test Set Evaluation Results		
Metric	**Value**	**Interpretation**
Gini	0.547	Fair (0.5-0.7: Moderate)
Weighted Kappa	0.497	Fair (0.4-0.6: Moderate)
Weighted MAE	0.815	Poor (>0.5)
ORC	0.665	Fair (70-84%)

📎 Conclusion on the model: The model performs significantly better then Ordinal Logistic Regression die to relaxation of the proportionality of odds assumption on several predictors. Metrics of model evaluation show fair results. The difference of Gini coefficient between the train and test sets are insignificant, indicating absence of overfitting.

Light Gradient Boosting Machine

A high-performance gradient boosting framework optimized for speed and efficiency, using histogram-based algorithms and leaf-wise growth. Excels with large datasets, handling non-linear relationships automatically while offering GPU support and categorical feature handling. Requires careful tuning but delivers strong predictive accuracy with computational efficiency.

Model Assumptions

1. **No strict linearity assumption**: Handles non-linear relationships.

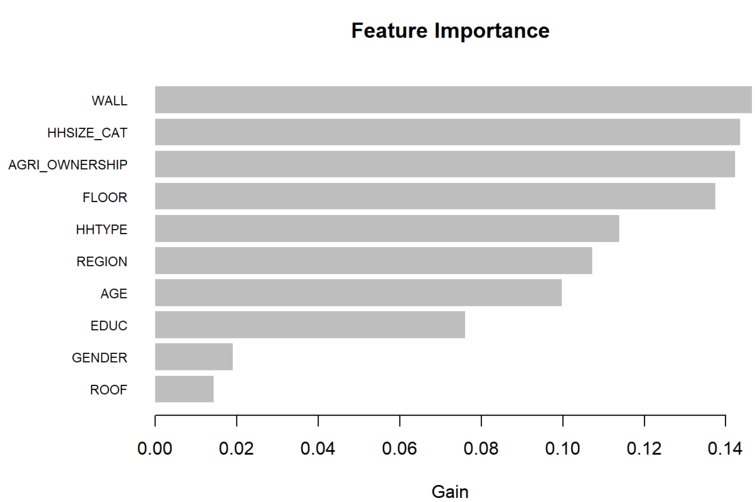
2. **Preprocess categoricals**: Use `categorical_feature` or one-hot encoding.
3. **Hyperparameter tuning**: Critical for performance (e.g., `num_leaves`, `learning_rate`).
4. **Robust to outliers**: But extreme values may affect splits.

Hyperparameter Tuning

The following results were obtained from Random Search hyperparameter tuning:

```
lgb_params <- list(
  num_leaves = 73,
  learning_rate = 0.1603422,
  feature_fraction = 0.8519221,
  min_data_in_leaf = 20,
  lambda_l1 = 2.903614,
  lambda_l2 = 1.354935,
  max_depth = 12,
  min_gain_to_split = 0.005324554,
  bagging_fraction = 0.97738,
  bagging_freq = 3,
  objective = "multiclass",
  metric = "multi_logloss",
  num_class = 5,
  min_split_gain = 0.01,
  feature_pre_filter = FALSE,
  verbose = 1
)
```

Feature Importance



Feature Interaction Table

```
           Parent          Child   sumGain frequency
           <char>         <char>     <num>     <int>
 1:          WALL          FLOOR 825052.5453       271
 2:         FLOOR    HHSIZE_CAT 796903.7084       382
 3:        HHTYPE AGRI_OWNERSHIP 791467.2841       101
 4:          EDUC            AGE 712466.4063       830
 5: AGRI_OWNERSHIP    HHSIZE_CAT 659599.6116       158
 6:         FLOOR           WALL 573482.6579       320
 7:        HHTYPE    HHSIZE_CAT 531039.8572       371
 8:        GENDER            AGE 494531.4966       411
 9:        HHTYPE          FLOOR 474359.7659       271
10:          WALL         HHTYPE 470625.8279       249
11:    HHSIZE_CAT AGRI_OWNERSHIP 457892.8371       122
12:           AGE           EDUC 453433.4267       690
13:          WALL     HHSIZE_CAT 449109.9422       322
14:    HHSIZE_CAT          FLOOR 436838.3284       367
15: AGRI_OWNERSHIP           WALL 416599.2166       145
16:    HHSIZE_CAT         HHTYPE 342684.7520       356
17:         FLOOR         HHTYPE 342495.2700       317
```

Model Evaluation

→ Train Set

Metric	Value
Multi-class area under the curve	0.9584
Gini	0.91677

→ Test Set

Model Performance Metrics

Test Set Evaluation Results

Metric	**Value**	**Interpretation**
Gini	0.917	**Excellent (≥0.7: Strong discrimination)**
Weighted Kappa	0.798	**Excellent (≥0.6: Substantial agreement)**
Accuracy	0.792	**Fair (70-79%)**
Precision	0.764	**Excellent (≥75%)**
Recall	0.773	**Fair (60-79%)**
F1 Score	0.768	**Excellent (≥0.75)**

ORC	0.870	Excellent (≥85%)

📎 Conclusion on the model: The Light GBM performance is excellent.

CatBoost Model

An advanced gradient boosting implementation featuring native handling of categorical variables through ordered boosting and innovative approaches to missing data. Particularly robust for heterogeneous datasets. Shows high quality results in a moderate speed while not requiring preprocessing of categorical variables

Model Assumptions

1. **Handles categoricals natively**: No need for one-hot encoding.
2. **Requires GPU for large data**: Optimal performance with GPU acceleration.
3. **Hyperparameter sensitivity**: Tune `depth`, `iterations`, and `learning_rate`.
4. **Automatic handling of missing values**: But imputation may still help.

Feature Independence Test

CatBoost, like most gradient boosting algorithms, does not require features to be statistically independent. It can model interactions and dependencies between features through the construction of decision trees. However, it is robust and effective even when features are independent.

```
## # A tibble: 55 × 6
##    Variable1 Variable2      ChiSq    df p_value Warning
##    <chr>     <chr>          <dbl> <int>   <dbl> <chr>
##  1 WEALTH    AGE          124259.    28 0       OK
##  2 WEALTH    GENDER        16294.     4 0       OK
##  3 WEALTH    EDUC         244927.    16 0       OK
##  4 WEALTH    HHTYPE      1251902.    28 0       OK
##  5 WEALTH    HHSIZE_CAT  1343130.    20 0       OK
##  6 WEALTH    AGRI_OWNERSHIP 2117631.  4 0       OK
##  7 WEALTH    FLOOR       1156970.     8 0       OK
##  8 WEALTH    WALL        1504440.     8 0       OK
##  9 WEALTH    ROOF         221873.     8 0       OK
## 10 WEALTH    REGION      3355114.    16 0       OK
## 11 AGE       GENDER        28498.     7 0       OK
## 12 AGE       EDUC        7686730.    28 0       OK
## 13 AGE       HHTYPE      1196934.    49 0       OK
## 14 AGE       HHSIZE_CAT   731593.    35 0       OK
```

```
## 15 AGE          AGRI_OWNERSHIP     7797.     7 0         OK
## 16 AGE          FLOOR            54476.    14 0         OK
## 17 AGE          WALL             52964.    14 0         OK
## 18 AGE          ROOF             31848.    14 0         OK
## 19 AGE          REGION           83454.    28 0         OK
## 20 GENDER       EDUC            227923.     4 0         OK
## 21 GENDER       HHTYPE           67903.     7 0         OK
## 22 GENDER       HHSIZE_CAT       22148.     5 0         OK
## 23 GENDER       AGRI_OWNERSHIP    5926.     1 0         OK
## 24 GENDER       FLOOR             5738.     2 0         OK
## 25 GENDER       WALL              5736.     2 0         OK
## 26 GENDER       REGION            6278.     4 0         OK
## 27 EDUC         HHTYPE          360103.    28 0         OK
## 28 EDUC         HHSIZE_CAT      466861.    20 0         OK
## 29 EDUC         AGRI_OWNERSHIP   41507.     4 0         OK
## 30 EDUC         FLOOR           150129.     8 0         OK
## 31 EDUC         WALL            161483.     8 0         OK
## 32 EDUC         ROOF             35670.     8 0         OK
## 33 EDUC         REGION          263861.    16 0         OK
## 34 HHTYPE       HHSIZE_CAT    19763144.    35 0         OK
## 35 HHTYPE       AGRI_OWNERSHIP  170289.     7 0         OK
## 36 HHTYPE       FLOOR           189676.    14 0         OK
## 37 HHTYPE       WALL            378964.    14 0         OK
## 38 HHTYPE       ROOF             88110.    14 0         OK
## 39 HHTYPE       REGION          453920.    28 0         OK
## 40 HHSIZE_CAT   AGRI_OWNERSHIP  223841.     5 0         OK
## 41 HHSIZE_CAT   FLOOR           431892.    10 0         OK
## 42 HHSIZE_CAT   WALL            355496.    10 0         OK
## 43 HHSIZE_CAT   ROOF            215479.    10 0         OK
## 44 HHSIZE_CAT   REGION          910990.    20 0         OK
## 45 AGRI_OWNERSHIP FLOOR          20100.     2 0         OK
## 46 AGRI_OWNERSHIP WALL          114080.     2 0         OK
## 47 AGRI_OWNERSHIP ROOF           71304.     2 0         OK
## 48 AGRI_OWNERSHIP REGION        897582.     4 0         OK
## 49 FLOOR        WALL            249740.     4 0         OK
## 50 FLOOR        ROOF             62535.     4 0         OK
## 51 FLOOR        REGION         1685502.     8 0         OK
## 52 WALL         ROOF            593653.     4 0         OK
## 53 WALL         REGION         2363488.     8 0         OK
## 54 ROOF         REGION         1148568.     8 0         OK
## 55 GENDER       ROOF              1061.     2 4.13e-231 OK
```

From the above Chi-Square Test we can conclude that there is no significance dependence among the chosen predictors.

Catboost Feature Importance

To understand how each feature contributes to the model's prediction let's compute Feature Importance.

To get the following table, CatBoost traverses trees, recording how much each feature contributes to prediction shifts.It averages over all trees in the ensemble. The result is a vector of importances (same length as number of features), often normalized so the sum is 100.

```
##                    [,1]
## AGE            8.442758
## GENDER         3.263070
## EDUC           8.034356
## HHTYPE        11.754859
## HHSIZE_CAT    15.877825
## AGRI_OWNERSHIP 16.961526
## FLOOR         15.599579
## WALL          13.335130
## ROOF           1.723804
## REGION         5.007093
```

Model Evaluation

Metric	Value
Gini (Train)	0.91578
Gini (Test)	0.9154
Weighted Kappa (Test)	0.7965
Accuracy	0.79137

📎 Conclusion on the model: Catboost performs excellent results. Time of compilation is long: 9 hours.

Comparison of the models

Metric	OLR	GLR	Light GBM	Catboost
Gini (Train)	0.483	0.548	0.916	0.915
Gini (Test)	0.000441979905060563	0.547	0.917	0.915
Weighted Kappa (Test)	0.0007372	0.497	0.798	0.796
ORC (Test)	0.00036373355246857	0.665	0.870	
Accuracy	0.199767946450715	0.469	0.792	0.791

Precision	0.287673601462297	0.179	0.764	
Recall	0.202291054206273	0.449	0.773	
F1 Score	0.237542832667047	0.256	0.768	

Conclusion

Based on the evaluation metrics, **LightGBM** and **CatBoost** significantly outperform **Ordinal Logistic Regression (OLR)** and **Generalized Linear Regression (GLR)** across all metrics. Both gradient boosting models achieve the highest values in Gini (Train and Test), Accuracy, Precision, Recall, and F1 Score, with LightGBM showing a slight edge overall. In contrast, OLR and GLR exhibit poor generalization, with extremely low Gini (Test) and near-zero Weighted Kappa and ORC, indicating weak predictive power. Among traditional models, GLR performs better than OLR but still falls short of acceptable thresholds. Therefore, **LightGBM** and **CatBoost** are the most suitable models for deployment in this context.