Data Science CS/CE 457-464

ANALYZING 54 YEARS OF PAKISTAN'S DEMOCRATIC JOURNEY

Data-Driven Insights from General Elections (1970-2024)

Members :

Hunain Abbas

Hammad Malik

Hayyan Khan

Abdul Wasay

Research Questions

1.How have regional party strongholds shifted over 54 years, especially PTI's rise in urban areas? (Done in Previous Presentation)

2. Can we forecast future turnout and identify swing constituencies using ML models? We will answer this in this presentation.

3. What socioeconomic factors (literacy, poverty, urbanization) influence voter turnout and electoral competitiveness across Pakistan?

Objectives and Problem Statement

Why Forecast turnout?

- Voter turnout is a crucial indicator of democratic engagement
- Low turnout can signal political apathy or disenfranchisement
- High turnout often indicates competitive, consequential elections

Why Study Socioeconomic Factors?

• Literacy, poverty, and urbanization strongly influence voter turnout

• These factors help explain why some constituencies participate more than others

• Combining socioeconomic data with election data improves model interpretability

• Helps understand structural inequalities behind democratic participation

Why Identify Swing Constituencies?

- Swing constituencies are politically unstable and can flip between parties
- These are battlegrounds where elections are won or lost
- Understanding what makes a constituency "swing" helps predict electoral outcomes

Our Approach

We use machine learning to predict both continuous (turnout rates) and categorical (swing/stable) outcomes based on historical electoral patterns. We also integrate socioeconomic indicators (literacy, urbanization, poverty) to understand how structural factors shape voter turnout.

Data Preparation & Feature Engineering

Dataset: **24,585** constituency-level records
across 11 elections (1970-2024)

These are the key features that we created :

Historical Patterns:

- Previous election turnout
- Turnout trends (change over time)
- Vote margin from last election

Demographic Context:

- Registered voters (log-transformed)
- Province-level patterns

Competitive Metrics

- Herfindahl-Hirschman Index (HHI) - measures party concentration
- Winner's vote share
- Number of competitive parties (>10% vote share)

Additional Engineered Features:

- Lag1_Turnout & Lag2_Turnout
- Province/District target encodings
- Urban–Rural classification
- Literacy, urbanization, poverty indicators

Data Quality: Handled missing values, removed outliers, standardized features

Understanding Swing Constituencies

What Makes a Constituency "Swing"?

- A constituency is "swing" when the victory margin is less than 10% of total votes, this indicates close competition where outcomes are unpredictable

Swing Threshold

- If the victory margin is less than 10% of total votes, we classify it as a swing constituency

Supporting Metrics:

- We also calculate HHI (Herfindahl-Hirschman Index) to measure party concentration
- HHI ranges from 0 (highly fragmented) to 1 (one party dominates)
- HHI is used as a predictive feature in our model

Example: A close race where the winner gets 40% and runner-up gets 35% (5% margin) would be swing. HHI helps us understand competitiveness, but swing classification uses the 10% margin threshold.

Why Include Socioeconomic Factors?

Motivation:

- Electoral behavior is not driven by political factors alone
- Socioeconomic inequality directly impacts voter participation
- Understanding literacy, poverty, and urbanization helps explain why turnout varies
- Strengthens the interpretability of ML predictions beyond numerical patterns

Key Socioeconomic Variables Used:

- **Literacy Rate:** Higher education levels → higher political engagement
- **Urbanization Rate:** Urban constituencies show more competition and higher turnout
- **Poverty Index (MPI):** Higher poverty → lower turnout due to structural exclusion
- **District Development Data**: Extracted from district profiles and merged with election data

Data Source & Extraction Method:

- Extracted from government district profiles (DOCX format)
- Parsed and merged into the main dataset at the district level
- Standardized and encoded for ML modeling

Model Selection & Training

Why Gradient Boosting?

Model Choice Justification:

- Sequential Learning: Builds trees that correct previous errors
- Handles Non-linearity: Captures complex political patterns
- Feature Interactions: Automatically learns relationships between demographics, history, and outcomes
- Robust to Outliers: Works well with diverse constituency sizes

Training Strategy:

- GroupKFold Cross-Validation: Ensures constituencies stay together (prevents data leakage)
- 5-fold validation: Tests model on unseen constituencies
- Prevents Overfitting: Validates that patterns generalize beyond training data

Model Parameters:

- 250 estimators (trees) per model
- Maximum depth of 3 levels
- Learning rate: 0.05

Model Selection & Training for SocioEconomic

Ordinary Least Squares (Reduced)
Model Choice Justification:

- **Multicollinearity Fix:** Detected severe overlap between Literacy and Income (VIF > 50). Removing Income (VIF < 3) stabilized the model.
- **Interpretability:** OLS offers direct, explainable coefficients

Configuration:
OLS: Included intercept term (add_constant) for baseline measurement.
Regularization Baselines: Tested Ridge ($\alpha=1.0$) and Lasso ($\alpha=0.5$) for comparison but found no significant performance gain over the cleaned OLS model.

Training Strategy:

- **Stratified Split:** 80/20 train-test split stratified by Province to ensure equal regional representation.
- **Preprocessing:** Applied StandardScaler (Z-score) to normalize varying feature scales
- **Validation:** Employed 5-Fold Cross-Validation to verify results were robust and not specific to one data split.

Model Parameters:

- Parameters: Constant added; Ridge ($\alpha=1.0$) & Lasso ($\alpha=0.5$) used for comparison.
- Data Split: 80% Train / 20% Test, stratified by Province to ensure regional balance.
- Verification: Validated via 5-Fold CV to ensure model stability.

Turnout Forecasting Results

Model Performance: Predicting Voter Turnout

Metric	Score	Interpretation
R^2 Score	0.564	Model explains **56.4%** of turnout variation
MAE	0.77%	Average error is less than **1 percentage point**

Turnout Forecasting Results

What This Means:

- Our model can predict constituency-level turnout with ~1% accuracy
- Significantly better than baseline (always predicting average)
- Reliable enough for strategic electoral planning

Key Insight:

- Historical turnout is the strongest predictor, followed by constituency size and previous margins of victory.
- Limitations: Cannot predict unprecedented events (e.g., major political crises, new voting laws)

Swing Constituency Detection Results

Classification Performance:

Metric	Score	Interpretation
ROC-AUC	0.703	70% probability of correctly ranking swing vs stable
F1-Score	50.20%	Balanced precision-recall tradeoff

Swing Constituency Detection Results

Model Interpretation:

- ROC-AUC of 0.703: Model performs significantly better than random guessing (0.5)
- Model is conservative in predicting swings to avoid false alarms
- Better at identifying stable constituencies than detecting all swings

Challenge:

- Swing constituencies are inherently unpredictable - they're defined by change!

Trade-off:

- We prioritized avoiding false positives (predicting stable as swing) over catching every swing

Swing Constituency Detection Results

Top Swing Constituencies (Next Election)

Rank	Constituency (Prediction for Next Election)	Predicted Swing Probability	Strategic Insight
1	NA-264 - Quetta 3	0.919	Extremely high likelihood of a close race/flip. Top battleground.
2	NA-262 - Quetta 1	0.919	Also highly likely to be a swing seat. Quetta/Balochistan shows high potential volatility.
3	NA-94 - Chiniot 2	0.719	Strong probability of high competition.
4	NA-97 - Faisalabad 3	0.577	Moderate probability, warrants attention.

Justifying Model Performance

MAE (Mean Absolute Error) — 0.77%

R^2 (Coefficient of Determination) — 0.564

In social science, where human behavior is complex and partly unobservable, R^2 values above 0.10 are often acceptable, and 0.3–0.5 are considered moderate to substantial fits (Ozili, 2023).
A score of 0.564 is therefore strong, domain-appropriate, and robust for electoral forecasting.

Justifying Model Performance

ROC–AUC — 0.703

An AUC of 0.5 reflects performance equivalent to random chance. In contrast, an AUC within the 0.70–0.80 range is commonly interpreted as indicating good discriminatory power (Keylabs, 2024). Thus, our score demonstrates that the model offers meaningful predictive value for prioritization.

F1-Score — 50.20%

The F1-Score is the metric of choice for imbalanced datasets because the harmonic mean heavily penalizes low values in either precision or recall, providing a more realistic assessment of performance on the minority class than accuracy. (KeyLabs, 2024)

Socioeconomic Analysis Results

Predicting Turnout via Socioeconomic Factors

Metric	Score
Adjusted R^2	0.232
Literacy Impact (β)	0.34
Urbanization P-value	0.195

Swing Constituency Detection Results

Model Interpretation:

- The Literacy Lever: A 10% rise in literacy translates to approximately 3.4% higher turnout.
- The Urban Myth: Urbanization has zero impact ($p=0.19$) on voter participation once literacy is accounted for.
- The "Unexplained": The model explains 23% of the variance; the remaining 77% is driven by politics (mobilization, electables), not demographics.

Challenge:

- Multicollinearity Trap: Literacy and Wealth were correlated at 0.91, making it initially impossible to mathematically separate their effects.
- Ecological Fallacy: The analysis uses district-level averages, which risks masking individual voter choices (we track regions, not specific people).

Trade-off:

- Stability vs. Completeness: We dropped the Income variable to fix the VIF score, prioritizing stable coefficients over including every possible feature.
- Inference vs. Prediction: We selected simple OLS Regression over complex "Black Box" models to ensure we could explain exactly how factors impact turnout.

Limitations of our Study

What We Cannot Predict:

- Major political crises, party bans, or electoral irregularities
- New political movements (e.g., PTI's initial rise in 2013)
- Campaign effects: candidate quality, spending, media coverage
- External shocks: military interventions, constitutional changes

Missing Data:

- Constituency boundary changes over 54 years
- Social media influence and youth mobilization patterns

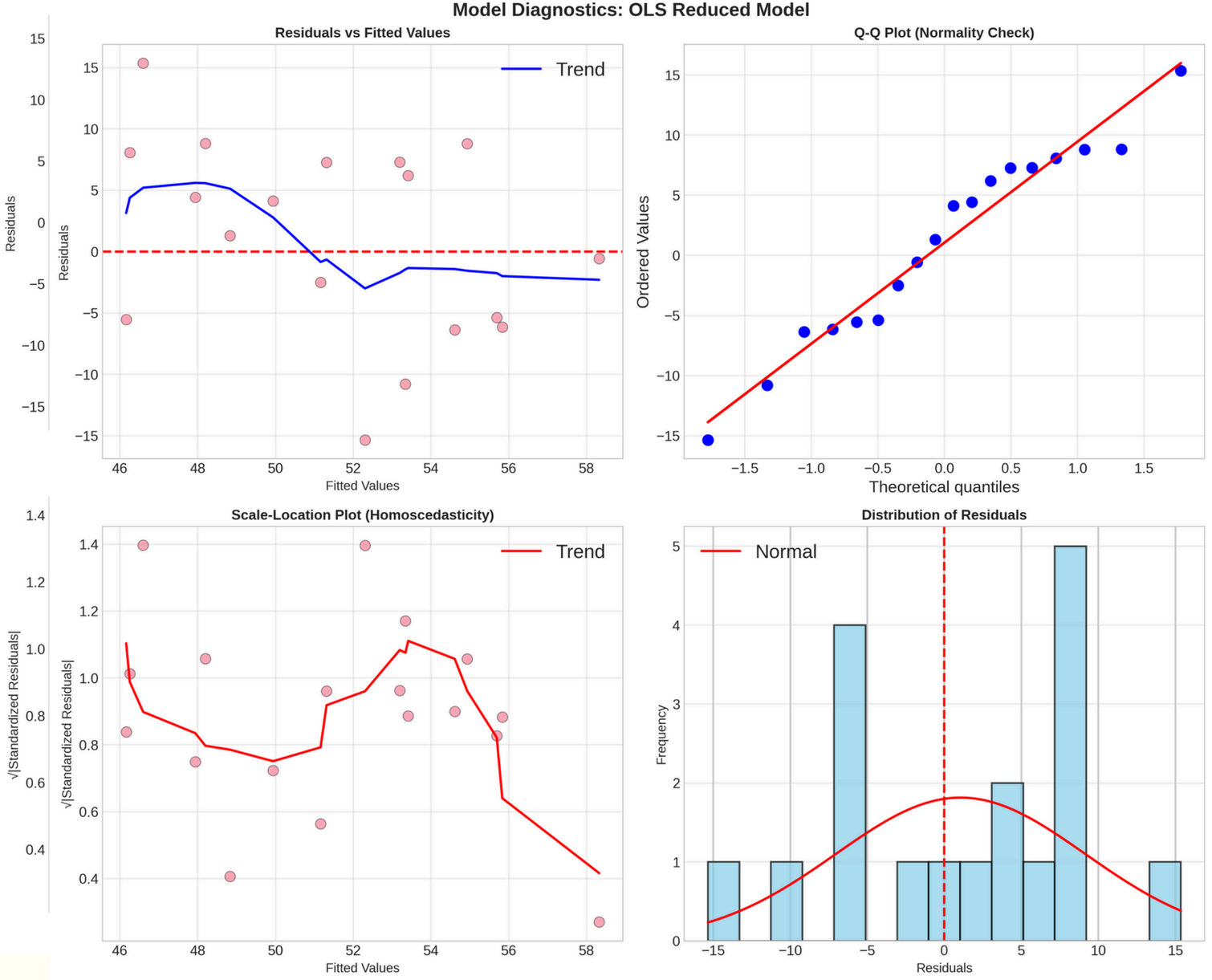
Model Trade-offs:

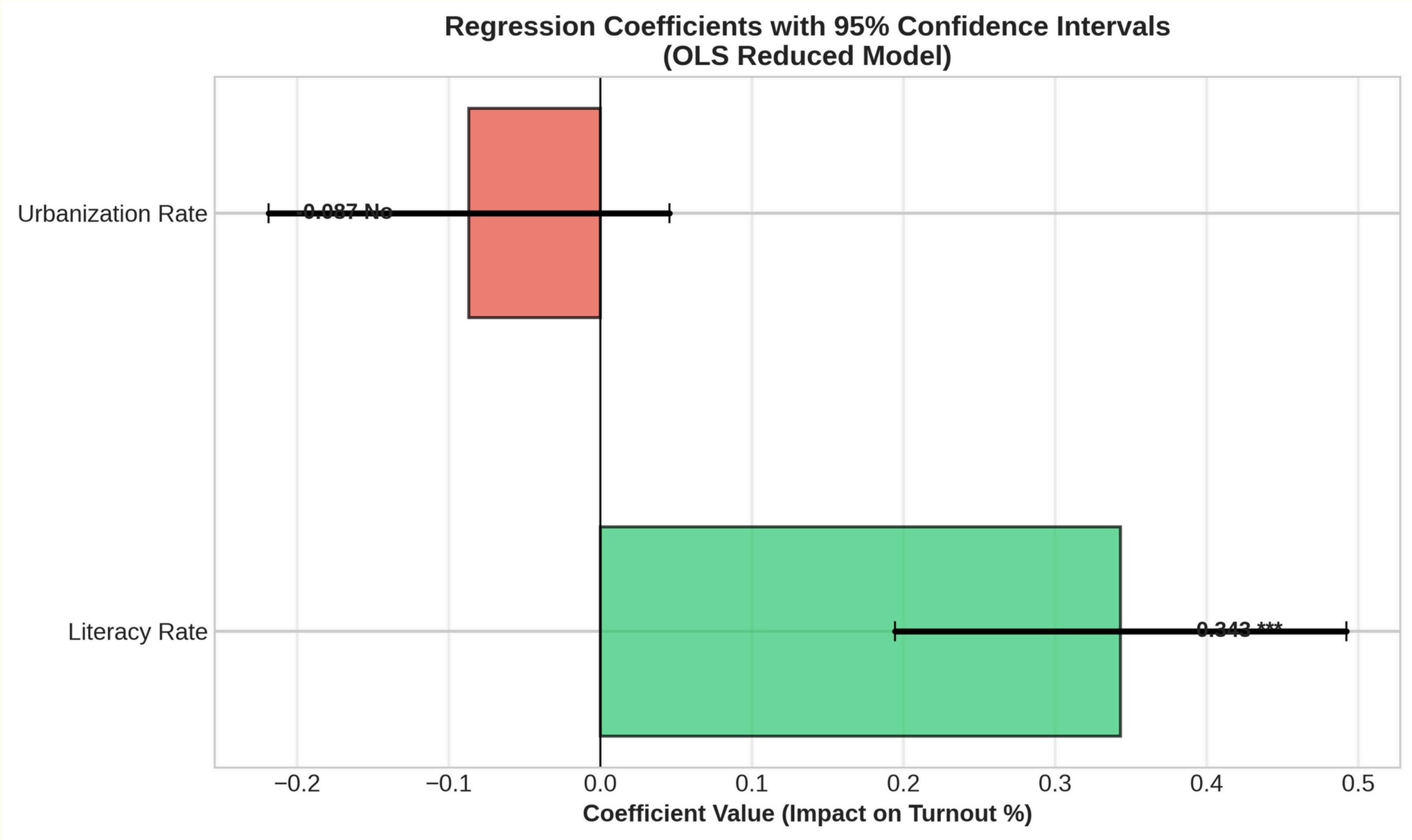
- Conservative on swing predictions (avoids false alarms)
- Assumes past patterns continue into the future
- Works best for stable political environments

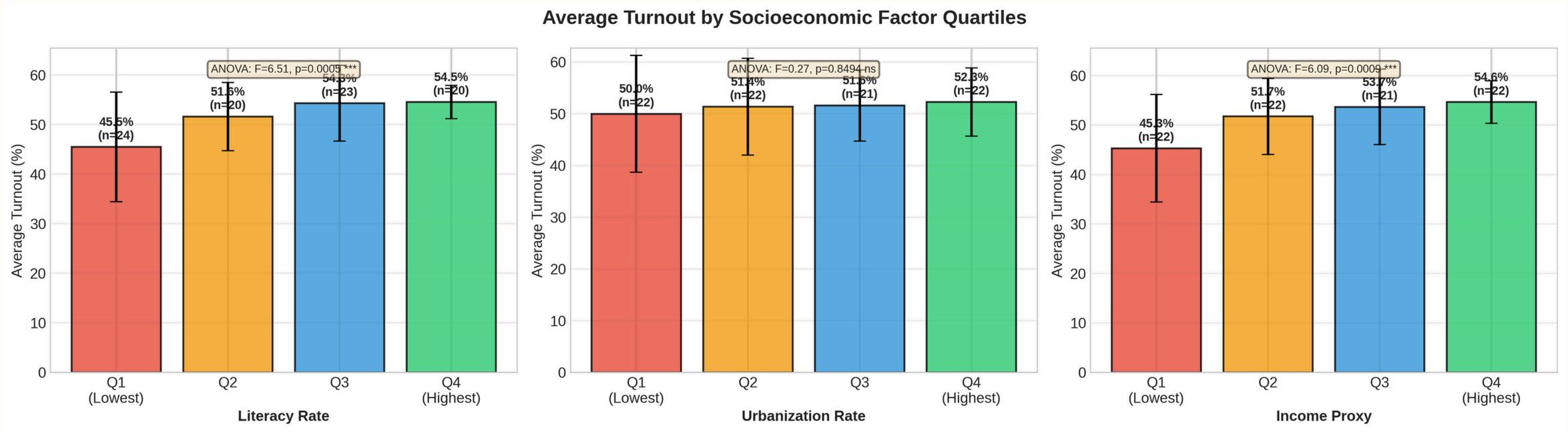
Graphs

Model Diagnostics: OLS Reduced Model

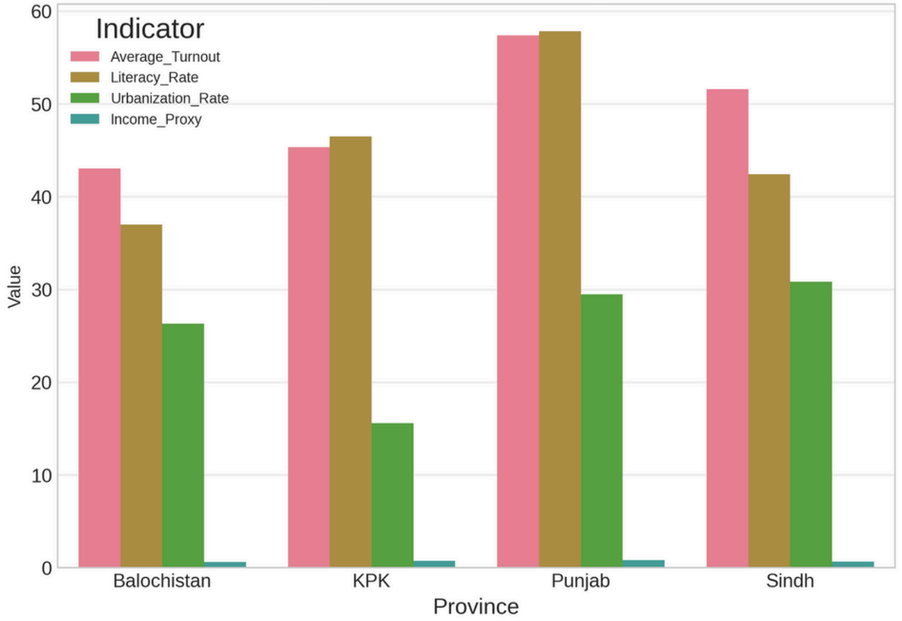
Model Diagnostics: OLS Reduced Model

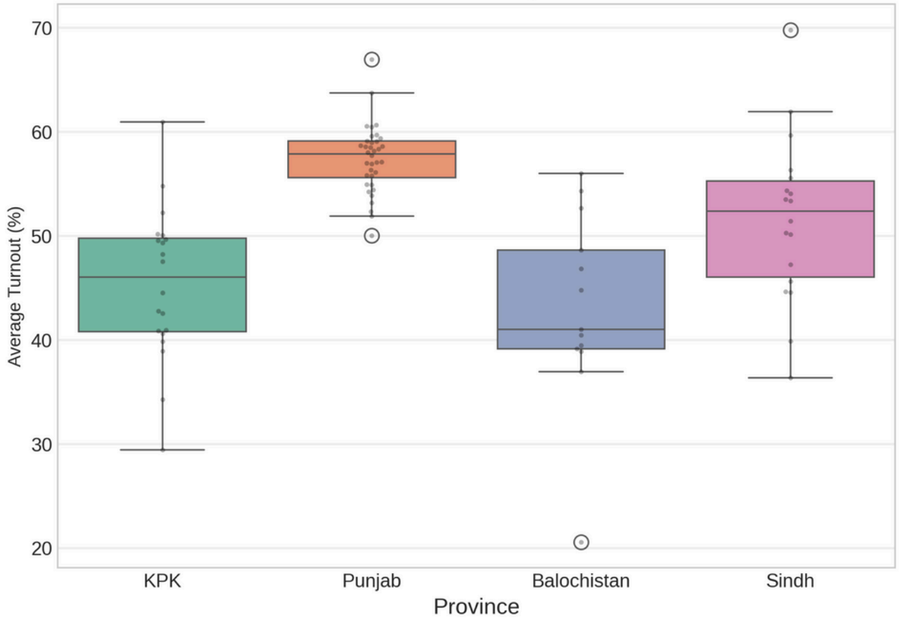


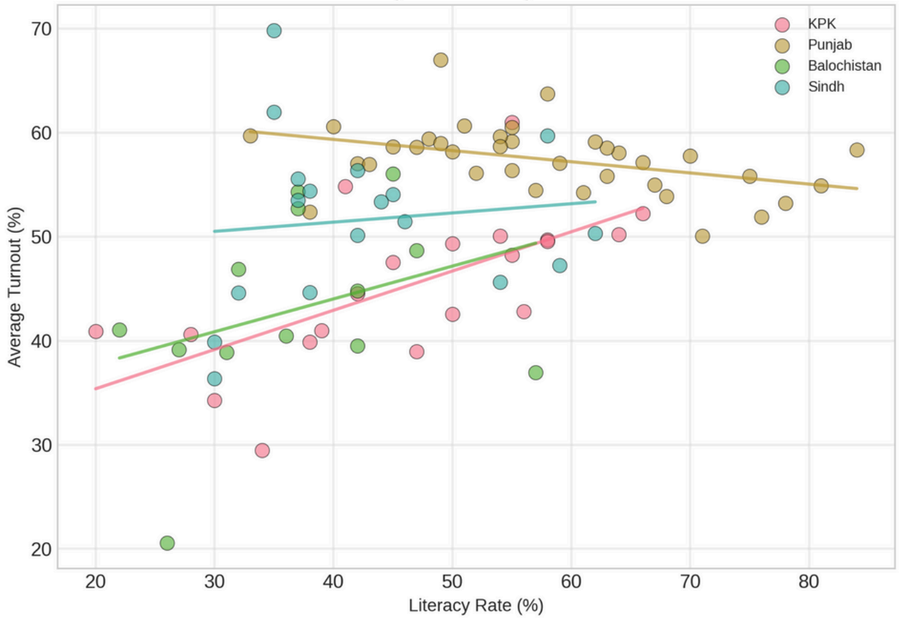

Regression Coefficients with 95% Confidence Intervals
(OLS Reduced Model)
Urbanization Rate
-0.087 ns
Literacy Rate
0.343 ***
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Coefficient Value (Impact on Turnout %)

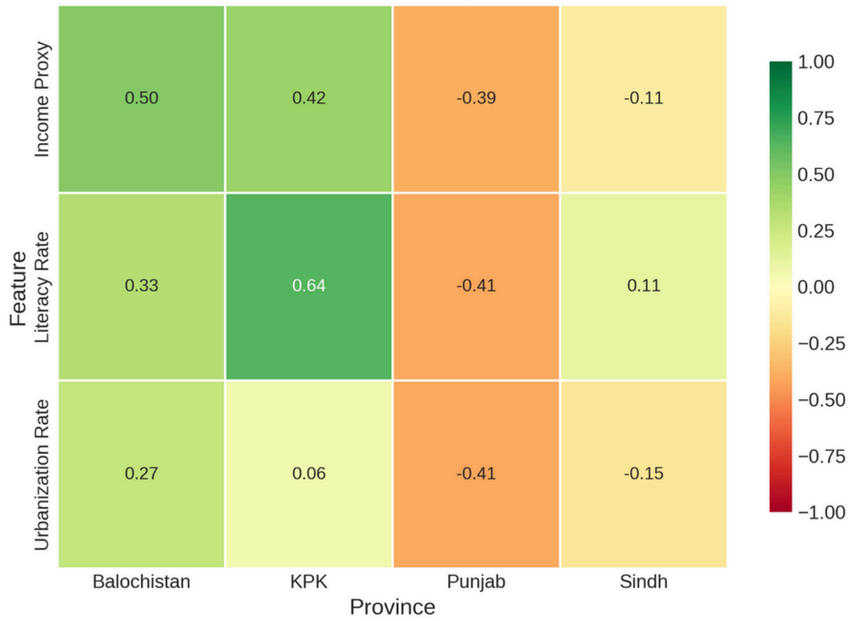

Average Turnout by Socioeconomic Factor Quartiles
ANOVA: F=6.51, p=0.0005 ***
45.5%
(n=24)
51.6%
(n=20)
54.5%
(n=23)
54.5%
(n=20)
Average Turnout (%)
Q1
(Lowest)
Q2
Q3
Q4
(Highest)
Literacy Rate
ANOVA: F=0.27, p=0.8494 ns
50.0%
(n=22)
51.4%
(n=22)
51.6%
(n=21)
52.3%
(n=22)
Urbanization Rate
ANOVA: F=6.09, p=0.0009 ***
45.3%
(n=22)
51.7%
(n=22)
53.7%
(n=21)
54.6%
(n=22)
Income Proxy

Provincial Analysis: Socioeconomic Factors and Turnout

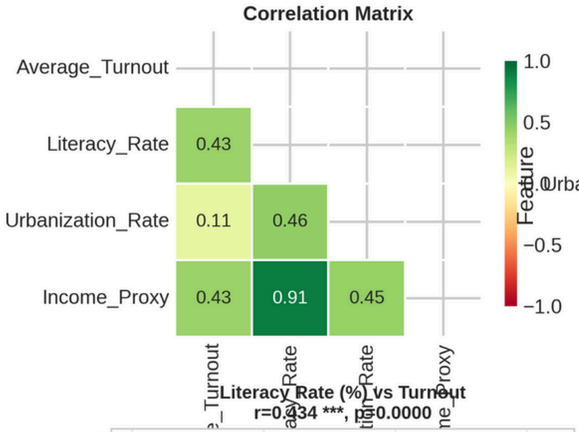


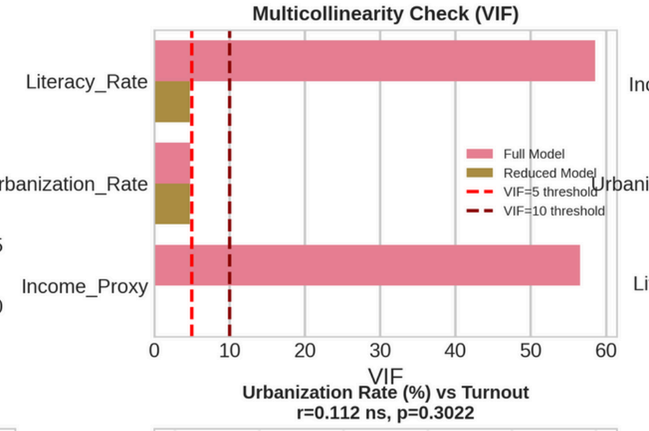


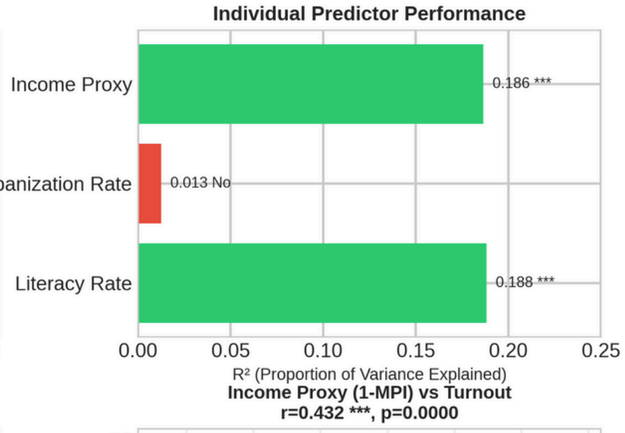


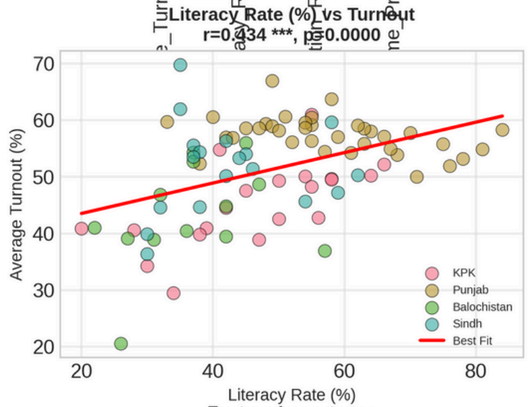


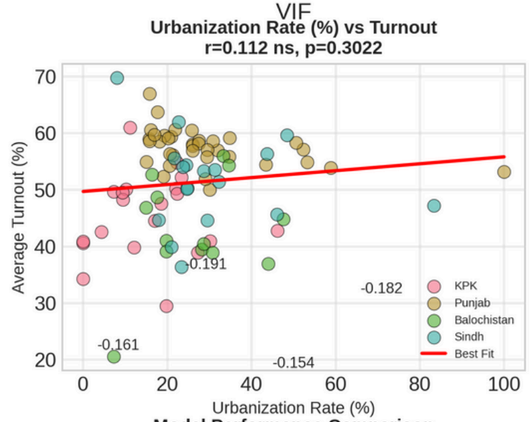
Which Socioeconomic Factors Best Predict Voter Participation?
Comprehensive Analysis of 2018 Pakistan Elections

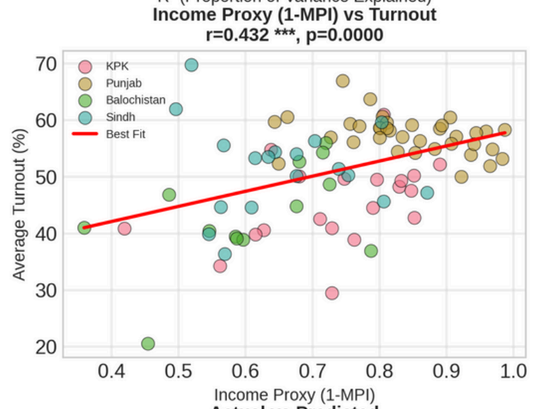


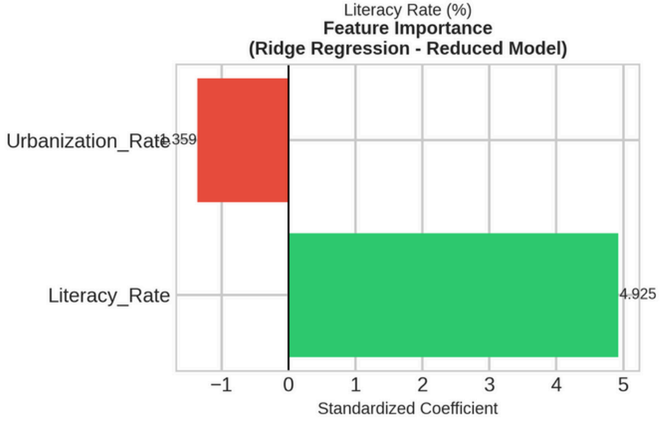


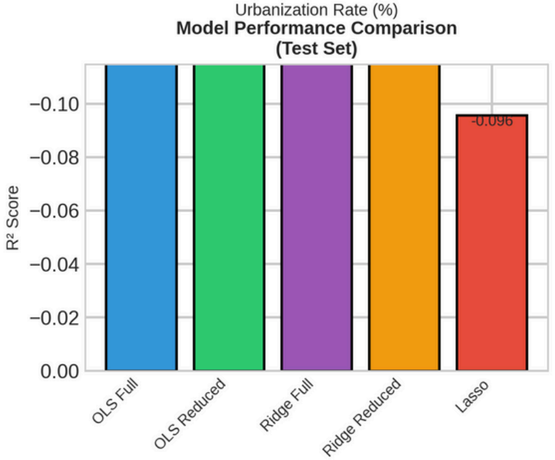


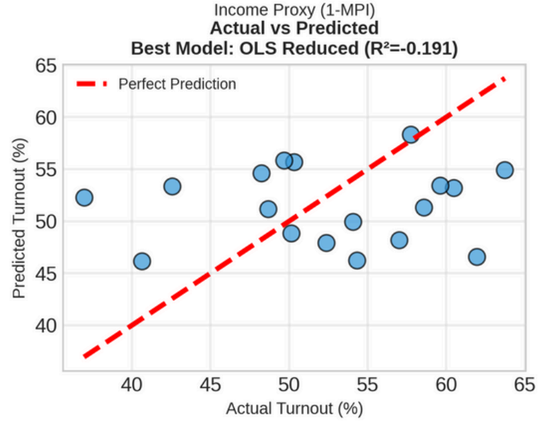












THE END