APPLIED MACHINE LEARNING IC

Unleashing the Power of AI

REGRESSION

Prof. Dr. Sinem Solak

Professor for AI & Technology
Program Director of BSc in AI and Sustainable Technologies

This Challenge is brought to you by

Sinem & Georg

Prof Dr Sinem Solak

🎓Professor for AI & Technology

📍 London, UK

Georg Bollweg

🎓 Challenge Tutor @ ToU

📍 Munich, Germany

🔍 Applied Mathematics @LMU

Reflection Session

Agenda

5’	Driving Question & Chat Discussion
20’	Review
25’	Reflections
10’	Q&A

"What surprised you most about turning a real-world sustainability problem into a working regression model and what would you do differently if you started over today?"

💬 Type your answer in the Chat!

Review

What is Regression? (Core Definition)

Key Message: "Predicting Continuous Values"

- Definition: Regression is a supervised learning technique for predicting continuous numerical outcomes
- Core Principle: Find the relationship between input features (X) and output target (y)
- Output: A continuous value (not a category)
- Goal: Minimize prediction error

Types of Regression Models

Model Type	Use Case	Complexity	Interpretability
Linear Regression	Linear relationships, baseline	Low	Very High
Polynomial Regression	Curved relationships	Medium	High
Ridge/Lasso Regression	Many features, regularization needed	Medium	High
Support Vector Regression	Non-linear patterns	High	Medium
Decision Tree Regression	Non-linear, interactions	Medium	High
Ensemble Methods (Random Forest, Gradient Boosting)	Complex patterns	High	Medium

Linear Regression Foundations

"The Workhorse Model"

- Formula:
  $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{□} x_{□}$

  β_0 = intercept

  $\beta_1, \beta_2, \ldots \beta_{□}$ = coefficients (slopes)

  $\hat{y}$ = predicted value

- Objective: Minimize Sum of Squared Residuals (SSR)
- Residual: difference between actual and predicted value ($y - \hat{y}$)
- Least Squares Method: Mathematical approach to find optimal coefficients

Key Assumptions of Linear Regression

"These Matter for Valid Results"

1. Linearity: Relationship between X and y is linear
2. Independence: Observations are independent (no autocorrelation)
3. Homoscedasticity: Constant variance of residuals across all X values
4. Normality: Residuals are normally distributed
5. No Multicollinearity: Features are not highly correlated with each other

Model Evaluation Metrics

Metric	Formula	Interpretation
MAE (Mean Absolute Error)	$(1/n)\Sigma\|y - \hat{y}\|$	Average absolute error in original units
MSE (Mean Squared Error)	$(1/n)\Sigma(y - \hat{y})^2$	Penalizes larger errors more heavily
RMSE (Root Mean Squared Error)	$\sqrt{}$MSE	Back to original units, interpretable
R^2 (Coefficient of Determination)	$1 - (SS\_res / SS\_tot)$	Proportion of variance explained (0 to 1)
Adjusted R^2	$1 - [(1-R^2)(n-1)/(n-p-1)]$	R^2 adjusted for number of features
MAE vs RMSE	Context-dependent	MAE: robust to outliers; RMSE: penalizes outliers

Overfitting vs. Underfitting

- Underfitting: Model too simple, misses true relationships
    - High bias, low variance
    - Poor performance on both training and test data

- Overfitting: Model too complex, memorizes noise
    - Low bias, high variance
    - Great on training data, poor on test data

- Sweet Spot: Model complexity balanced with generalization

Cross-Validation

"Rigorous Model Selection"

- Purpose: Estimate model performance on unseen data
- K-Fold Cross-Validation: Split data into k equal parts
  - Train on k-1 folds, test on 1 fold
  - Repeat k times, average results
- Typical Values: k=5 or k=10
- Advantage: Uses all data for both training and testing
- Disadvantage: Computationally expensive for large datasets

Data split into 5 folds:

Fold 1: [TRAIN] [TRAIN] [TRAIN] [TRAIN] [TEST]

Fold 2: [TRAIN] [TRAIN] [TRAIN] [TEST] [TRAIN]

Fold 3: [TRAIN] [TRAIN] [TEST] [TRAIN] [TRAIN]

Fold 4: [TRAIN] [TEST] [TRAIN] [TRAIN] [TRAIN]

Fold 5: [TEST] [TRAIN] [TRAIN] [TRAIN] [TRAIN]

Average scores across all folds

Feature Engineering for Regression

- Numerical Features:
  - Scaling/Normalization (StandardScaler, MinMaxScaler)
  - Polynomial features (capture non-linearity)
  - Interaction terms (e.g., length × width for area)
- Categorical Features:
  - One-hot encoding
  - Target encoding
  - Ordinal encoding (for ordered categories)
- Feature Selection:
  - Remove low-variance features
  - Remove highly correlated features
  - Use domain knowledge
- Handling Missing Data:
  - Deletion (if < 5% missing)
  - Mean/median imputation
  - Forward fill (time series)
  - Model-based imputation

Regularization Techniques

Technique	Penalty	Effect	When to Use
Ridge Regression (L2)	$\lambda \Sigma \beta^2$	Shrinks all coefficients proportionally	Many correlated features
Lasso Regression (L1)	$\lambda \Sigma \|\beta\|$	Pushes some coefficients to exactly zero	Feature selection needed
Elastic Net	$\lambda_1 \Sigma \beta^2 + \lambda_2 \Sigma \|\beta\|$	Combines Ridge and Lasso	Balanced regularization

Hyperparameter λ (lambda):
Controls regularization strength

- $\lambda = 0$: ordinary linear regression
- $\lambda \to \infty$: all coefficients $\to 0$
- Find optimal λ via cross-validation

Common Pitfalls & How to Avoid Them

Pitfall	Why It's Bad	How to Avoid
Using Test Data to Select Features	Leaks information, inflates performance	Feature selection on training data only
No Baseline Model	Can't judge if your complex model is worth it	Always build simple linear regression first
Ignoring Class Imbalance	(More relevant for classification, but matters for some regression)	Check distribution of target variable
Extreme Outliers Unchecked	Can drastically pull regression line	Visualize data, investigate outliers, consider robust regression

Common Pitfalls & How to Avoid Them

Not Scaling Before Regularization	Regularization weights features by magnitude	Always scale before Ridge/Lasso
Reporting Training Error Only	Overfitting goes undetected	Always report train AND test/validation metrics
Multicollinearity Ignored	Unstable coefficients, unreliable interpretations	Check correlations, use Ridge/Lasso, remove redundant features
Ignoring Temporal Structure	Violates independence assumption for time series	Use time series-specific models (ARIMA, Prophet) or lag features

Communicating Results to Stakeholders

- Avoid Technical Jargon: Say "predictions are accurate ±5 tonnes" not "RMSE = 5"
- Visualizations:
  - Actual vs. Predicted scatter plot (shows accuracy)
  - Residual plots (shows model reliability)
  - Feature importance / coefficient plot (shows drivers)
  - Prediction intervals (shows uncertainty)
- Key Messages:
  - What can the model predict?
  - What are the main drivers?
  - How confident are we? (quantify uncertainty)
  - What are limitations?

The Professional Learning Landscape

What You Actually Built vs. What You Thought You Were Building

```
              📊 YOUR ACTUAL ACHIEVEMENT SCOPE

What You Thought:              What You Actually Built:
┌──────────────────┐           ┌─────────────────────────┐
│ • Build a model  │           │ • Problem formulation   │
│ • Get good R²    │     →     │ • Data acquisition      │
│ • Submit report  │           │ • Quality assessment    │
└──────────────────┘           │ • Preprocessing system  │
                               │ • Feature engineering   │
                               │ • Model comparison      │
                               │ • Validation framework  │
                               │ • Stakeholder comms     │
                               │ • Ethical assessment    │
                               │ • Production planning   │
                               └─────────────────────────┘

            🏗️ INFRASTRUCTURE COMPONENTS: 85%
              🤖 ALGORITHM DEVELOPMENT: 15%
```

Problem Definition Evolution

From "Build AI for Sustainability" to Precise ML Questions

🎯 PROBLEM FORMULATION JOURNEY

INITIAL THINKING:		REFINED FORMULATION:
"Use AI to help environment"	→	"Predict building energy consumption from architectural parameters to optimize design decisions"
"Improve social outcomes"	→	"Estimate education funding needs based on demographic indicators for resource allocation"
"Reduce carbon emissions"	→	"Forecast transportation demand by route and time to optimize public transit scheduling"

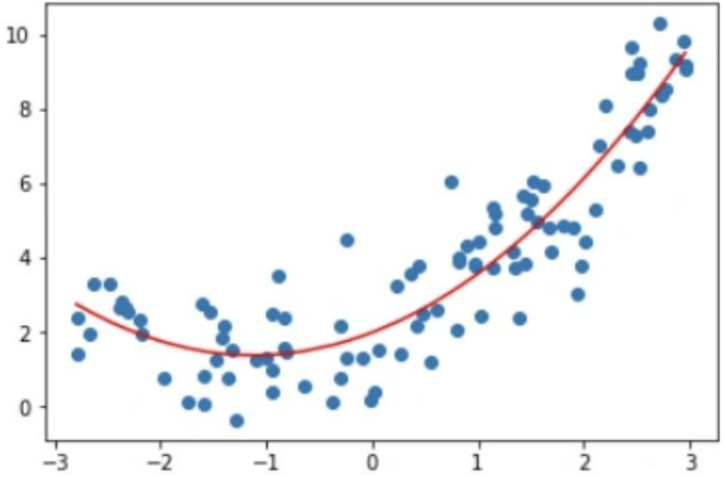
📋 PROBLEM QUALITY CHECKLIST:

✓ Specific, measurable target variable
✓ Available predictive features
✓ Actionable prediction timeline
✓ Clear stakeholder value proposition
✓ Ethical application boundaries

Practical Framework:

THE 5W+H TEST FOR ML PROBLEMS:

- WHO will use predictions?
- WHAT decision will they make?
- WHEN do they need predictions?
- WHERE will the model operate?
- WHY is prediction better than current approach?
- HOW will success be measured?

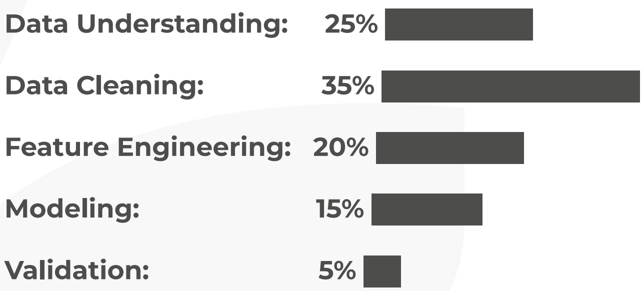


Data Reality vs. Expectations

DATA QUALITY REALITY CHECK

ACADEMIC DATASETS		YOUR REAL-WORLD DATA
✓ Complete	vs	⚠️ 15-30% missing
✓ Consistent		⚠️ Multiple formats
✓ Well-documented		⚠️ Sparse metadata
✓ Balanced		⚠️ Temporal gaps
✓ Large sample		⚠️ Geographic bias
✓ Clear labels		⚠️ Ambiguous targets

🕐 TIME ALLOCATION REALITY:

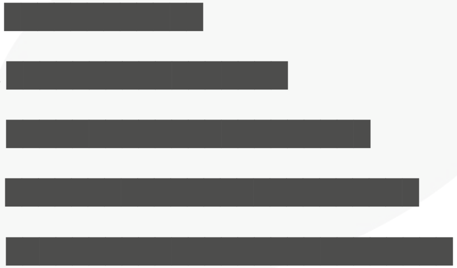


Feature Engineering Innovation

FEATURE ENGINEERING IMPACT ANALYSIS

PERFORMANCE GAINS BY FEATURE TYPE:

Raw Features Only:	$R^2 = 0.65$
+ Mathematical Transform:	$R^2 = 0.72$
+ Domain Calculations:	$R^2 = 0.78$
+ Temporal Features:	$R^2 = 0.83$
+ Interaction Terms:	$R^2 = 0.86$



🏆 MOST IMPACTFUL FEATURES ACROSS PROJECTS:

- Efficiency ratios (energy/output, cost/benefit)
- Normalized per-capita metrics (emissions/population)
- Temporal trends (3-month rolling averages)
- Geographic clustering indicators
- Interaction terms (income × education, size × density)

⚖️ EFFORT vs IMPACT ANALYSIS:

High Impact, Low Effort:	Domain ratios, log transforms
High Impact, High Effort:	Temporal aggregations, clustering
Low Impact, Low Effort:	Simple interactions, binning
Low Impact, High Effort:	Complex polynomial terms

Algorithm Selection Reality

Beyond "Best Cross-Validation Score"

🏆 **ALGORITHM SELECTION FRAMEWORK**

PERFORMANCE vs CONTEXT MATRIX:

	Interpretability	Performance	Deployment	Maintenance
Linear Regression	████████	██████	████████	████████
Ridge/Lasso	████████	███████	████████	████████
Random Forest	██████	████████	██████	██████
Gradient Boosting	████	█████████	████	████
Neural Networks	██	█████████	███	███

ACTUAL PERFORMANCE COMPARISON (Average across projects):

Algorithm	R²	RMSE	Training_Time	Prediction_Speed	Explainability
Linear	0.76	0.34	2 sec	<1ms	High
Ridge	0.78	0.32	3 sec	<1ms	High
Lasso	0.77	0.33	5 sec	<1ms	Medium-High
Random Forest	0.82	0.29	45 sec	2ms	Medium
XGBoost	0.85	0.26	180 sec	5ms	Low

Evaluation Framework Sophistication

Measuring What Matters, Not Just What's Easy

📈 EVALUATION EVOLUTION JOURNEY

BASIC APPROACH → **SOPHISTICATED FRAMEWORK**

BASIC APPROACH

- R^2 score
- Training data
- Point estimate

SOPHISTICATED FRAMEWORK

- Multiple complementary metrics (R^2, RMSE, MAE)
- Cross-validation
- Confidence intervals
- Residual analysis
- Subgroup performance
- Business impact metrics
- Stakeholder alignment

BUSINESS IMPACT TRANSLATION:

Statistical Metric	→	Business Language
$R^2 = 0.82$	→	"Explains 82% of outcome variation"
RMSE = $2,340	→	"Average prediction error $2,340"
MAE = $1,890	→	"Typical error $1,890 (robust to outliers)"
95% CI = ±$4,200	→	"Prediction uncertainty range ±$4,200"

STAKEHOLDER-ALIGNED METRICS BY DOMAIN:

Environmental: % Emissions reduction, Cost per ton CO_2 saved

Social: Equity index improvement, Coverage gap reduction

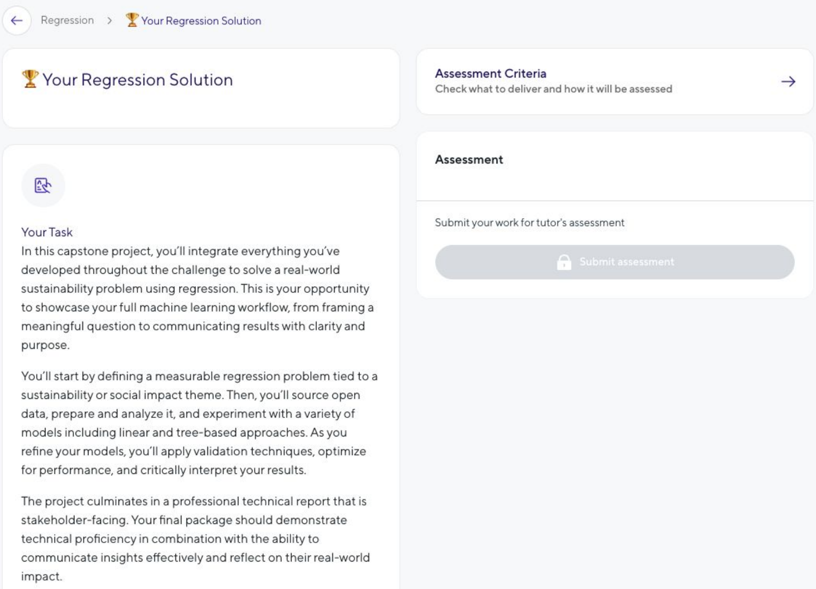
Economic: ROI per prediction, Cost avoidance achieved

Policy: Population affected, Geographic coverage achieved

Key Takeaways & Transferable Lessons

CORE ML PRINCIPLES

1. Define precise, impactful problems
2. Invest heavily in data quality & preprocessing
3. Engineer features that encode domain knowledge
4. Evaluate with diverse, business-aligned metrics
5. Embed ethics at every stage
6. Plan for production from day one
7. Communicate clearly to technical & non-technical audiences
8. Systematize your workflow for reproducibility & scalability



🏆 Your Regression Solution

Your Task

In this capstone project, you'll integrate everything you've developed throughout the challenge to solve a real-world sustainability problem using regression. This is your opportunity to showcase your full machine learning workflow, from framing a meaningful question to communicating results with clarity and purpose.

You'll start by defining a measurable regression problem tied to a sustainability or social impact theme. Then, you'll source open data, prepare and analyze it, and experiment with a variety of models including linear and tree-based approaches. As you refine your models, you'll apply validation techniques, optimize for performance, and critically interpret your results.

The project culminates in a professional technical report that is stakeholder-facing. Your final package should demonstrate technical proficiency in combination with the ability to communicate insights effectively and reflect on their real-world impact.

Assessment Criteria

Check what to deliver and how it will be assessed →

Assessment

Submit your work for tutor's assessment

WHAT IS THE PRIMARY GOAL OF REGRESSION IN MACHINE LEARNING?

EXPLAIN THE DIFFERENCE BETWEEN LINEAR AND NON-LINEAR REGRESSION

WHAT DOES THE TERM R^2 MEAN IN REGRESSION ANALYSIS?

HOW DO YOU INTERPRET THE COEFFICIENTS IN A LINEAR REGRESSION MODEL?

WHAT IS OVERFITTING IN REGRESSION, AND HOW CAN IT BE PREVENTED?

WHAT ROLE DOES REGULARIZATION PLAY IN REGRESSION MODELS?

- ## Additional Resources

Books:

- *An Introduction to Statistical Learning* by James, Witten, Hastie, Tibshirani (free PDF available)
- *The Hundred-Page Machine Learning Book* by Andriy Burkov

Online Resources:

- Scikit-learn documentation:
- https://scikit-learn.org/
- StatQuest with Josh Starmer (YouTube channel on regression)

Practice:

- Kaggle competitions with regression problems
- UCI Machine Learning Repository datasets

Sustainability-Specific:

- World Bank Open Data (climate, emissions)
- NASA Earth data (satellite climate data)

Sharing Your Unique Journey

Reflection

💬 Share your answer on the Miro Board!

Reflection

⏲️ **5’ Individual Work**
Answer the two questions on the Miro board (also considering your initial expectations from the Kick-Off Session).

⏲️ **5’ Peer Sharing**
Share your insights with your peers.

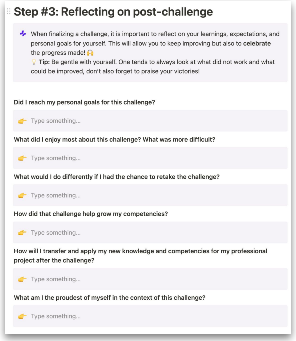
📕 **Learning Journal**
If you’d like, add your insights from this challenge to your learning journal.

Optional: get feedback from your peers for your learning journal. Sometimes outside perspectives are needed to help us really see how much we have grown.

✏️ How will I **transfer and apply** my new knowledge and competencies for my professional project during/after the challenge?

✏️ What have I actually **learned**? How will this support my **mission**?

✏️ What are my key insights from this challenge for my **learning journal**?



Step #3: Reflecting on post-challenge

When finalizing a challenge, it is important to reflect on your learnings, expectations, and personal goals for yourself. This will allow you to keep improving but also to **celebrate** the progress made! 🙌
💡 **Tip:** Be gentle with yourself. One tends to always look at what did not work and what could be improved, don’t also forget to praise your victories!

Did I reach my personal goals for this challenge?

👉 Type something...

What did I enjoy most about this challenge? What was more difficult?

👉 Type something...

What would I do differently if I had the chance to retake the challenge?

👉 Type something...

How did that challenge help grow my competencies?

👉 Type something...

How will I transfer and apply my new knowledge and competencies for my professional project after the challenge?

👉 Type something...

What am I the proudest of myself in the context of this challenge?

👉 Type something...

Wrap-Up

Key Dates of this Challenge

Submission 🎯
Sun,9 Nov, 6pm CET

Q&A