PREDICTING AI/ML SALARIES AND EVALUATING ERROR DIRECTION

Comparing ML and DL Algorithms

FERNANDO ISCAR
23- 10 - 2023

"THE AI-ML FIELDS HAVE BEEN GROWING EXPONENTIALLY FOR MORE THAN A DECADE

(Alekseeva et al., 2021)

"THE EASE OF AI IMPLEMENTATION HAS OPENED THE LABOR MARKET DOORS TO NON-EXPERTS

(Joshi, 2020, p. 247)

PROJECT MOTIVATION

Lack of research in salary predictions for AI-ML Fields

- **Often centered in broader range of job fields or specific markets**
- **Salary prediction for IT jobs is the most similar**
- **Bias exploration usually focused in the Gender pay gap**
  - **Usually common predictors scope (potential biases underexplored)**

SOCIAL RELEVANCE

Fair and equal employment practices are facilitated

Raise awareness for AI-ML job market.

SCIENTIFICAL RELEVANCE

Adressing the gap in AI-ML and lighting under-explored field of salary predictions with DL

Contribution to the body of research on salary estimation

RESEARCH QUESTIONS

Given an extensive developers survey dataset, how can AI/ML job salaries be predicted by first employing traditional predictors with a spectrum of ML/DL algorithms, and subsequently integrating less explored variables to study biases and potential discrimination nuances?

RQ 1

Utilizing traditional predictors like “Country”, “Job type”, “Education”, “Job title”, “Company size”, “Age”, “Experience”, and “Annual Salary”, how do various ML and DL algorithms such as MLR, LASSO, RIDGE, RF, XGBoost, and MLP perform in terms of prediction error and variance explained against a median-based baseline?

RQ 2

Selecting the best overall performer regression algorithm from the RQ1, how do additional features, including "Remote work", "Certifications", "Coding as a hobby", "Years Coding", "Gender", "Sexual orientation", "Ethnicity", "Physical disability", and "Mental disability", impact the prediction error and variance explained by the aforementioned model?”

RQ 3

By integrating attributes like "Gender", "Sexual orientation", "Ethnicity", "Physical disability", and "Mental disability" into the model, to what extent can potential patterns of discrimination be identified?

“HOW WAS IT DONE” (literature review)

- **Administrative and survey data**
- **Data cleaning (missing values, correct assymetry and inflation)**
- **Demographics, Education, Work type and Experience**

- **Regression or Classification problem**
- **Oftentimes a comparison of different models**

MLR LASSO RIDGE ENET DT RF XGBoost SVR

R2 RMSE MSE MAE

- **Several studies based on EDA**

(Wang., 2022).

(Jain et al., 2022).

(Brandwijk, 2021)

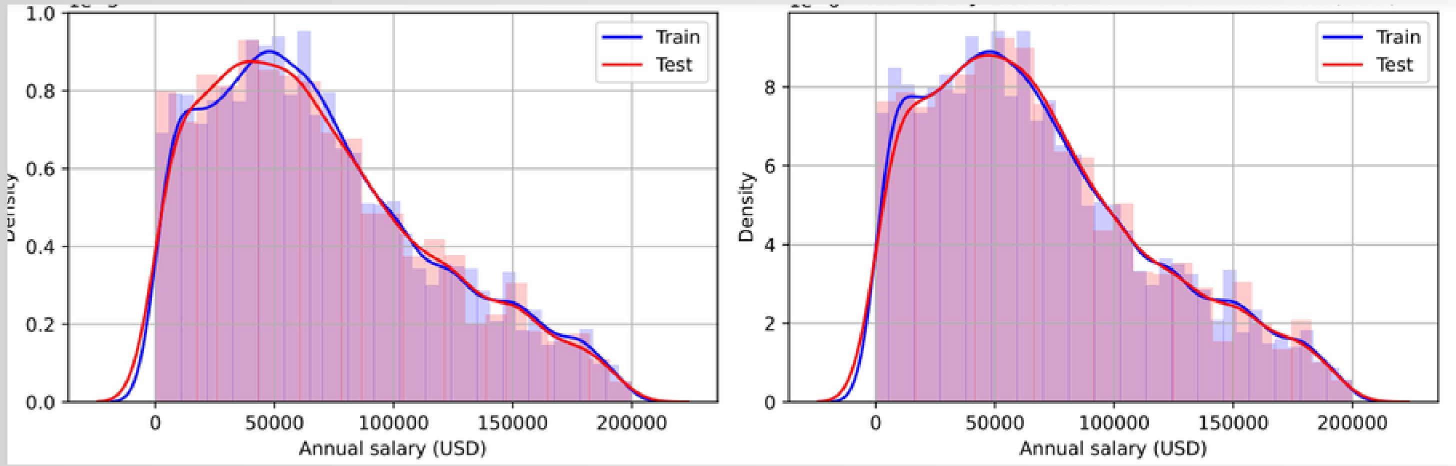
(Özer et al., 2022).

(Martin et al., 2018).

(Kablaoui & Salman, 2022)

(Matbouli & Alghamdi., 2022)

DATASET



Source

Stack Overflow
Annual Survey
(2022)

Original

81 Features

73.267 Rows

Missing values (>50%)
Class Imbalances (man-female)
Non - Symmetry (sqrt)

Processed

17 Features

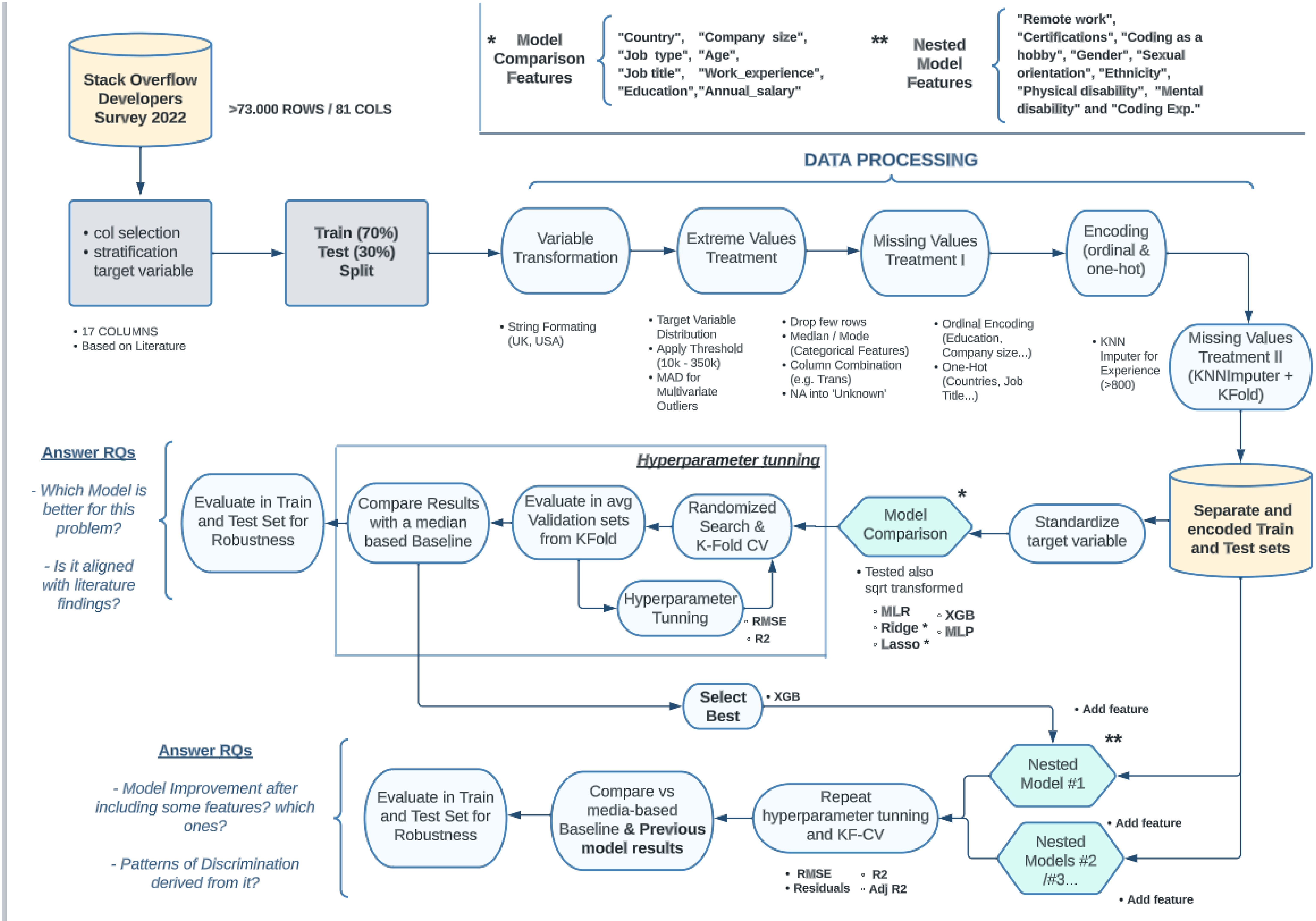
< 2.500 Rows

Filtered Job roles
Salary Thresholds
Dropping / imputing

METHODOLOGY

&

EVALUATION


* Model Comparison Features
"Country", "Company size", "Job type", "Age", "Job title", "Work_experience", "Education","Annual_salary"
** Nested Model Features
"Remote work", "Certifications", "Coding as a hobby", "Gender", "Sexual orientation", "Ethnicity", "Physical disability", "Mental disability" and "Coding Exp."
Stack Overflow Developers Survey 2022
>73.000 ROWS / 81 COLS
DATA PROCESSING
• col selection
• stratification target variable
• 17 COLUMNS
• Based on Literature
Train (70%) Test (30%) Split
Variable Transformation
• String Formating (UK, USA)
Extreme Values Treatment
• Target Variable Distribution
• Apply Threshold (10k - 350k)
• MAD for Multivariate Outliers
Missing Values Treatment I
• Drop few rows
• Median / Mode (Categorical Features)
• Column Combination (e.g. Trans)
• NA into 'Unknown'
Encoding (ordinal & one-hot)
• Ordinal Encoding (Education, Company size...)
• One-Hot (Countries, Job Title...)
• KNN Imputer for Experience (>800)
Missing Values Treatment II (KNNImputer + KFold)
Separate and encoded Train and Test sets
Standardize target variable
Model Comparison
*
• Tested also sqrt transformed
MLR
Ridge *
Lasso *
XGB
MLP
Hyperparameter tunning
Randomized Search & K-Fold CV
Hyperparameter Tunning
RMSE
R2
Evaluate in avg Validation sets from KFold
Compare Results with a median based Baseline
Evaluate in Train and Test Set for Robustness
Answer RQs
- Which Model is better for this problem?
- Is it aligned with literature findings?
Select Best
• XGB
• Add feature
**
Nested Model #1
• Add feature
Nested Models #2 /#3...
• Add feature
Repeat hyperparameter tunning and KF-CV
• RMSE
• Residuals
R2
Adj R2
Compare vs media-based Baseline & Previous model results
Evaluate in Train and Test Set for Robustness
Answer RQs
- Model Improvement after including some features? which ones?
- Patterns of Discrimination derived from it?

THANK YOU!