ReDI School of Digital Integration

ANALYTICS PROJECT

SALARY INSIGHTS: PREDICTING EARNINGS

LEARN FROM THE WORLD'S LARGEST COMMUNITY OF PROFESSIONAL SOFTWARE DEVELOPERS

Data Circle, Fall 2024, Berlin, Germany

Agenda

- Introduction
- Insights and trends in the developer community
- Machine learning models to predict annual salary
- Takeaways

STACK OVERFLOW
2023 DEVELOPER SURVEY

Data source: survey.stackoverflow.co

Contains locations, age, employment roles, education level, and annual salaries as well as preferred programming languages and years of professional coding

Developers Community

~90,000 developers

46% of respondents specify their salary

from 185 countries

80% of respondents have a Bachelor's degree or higher

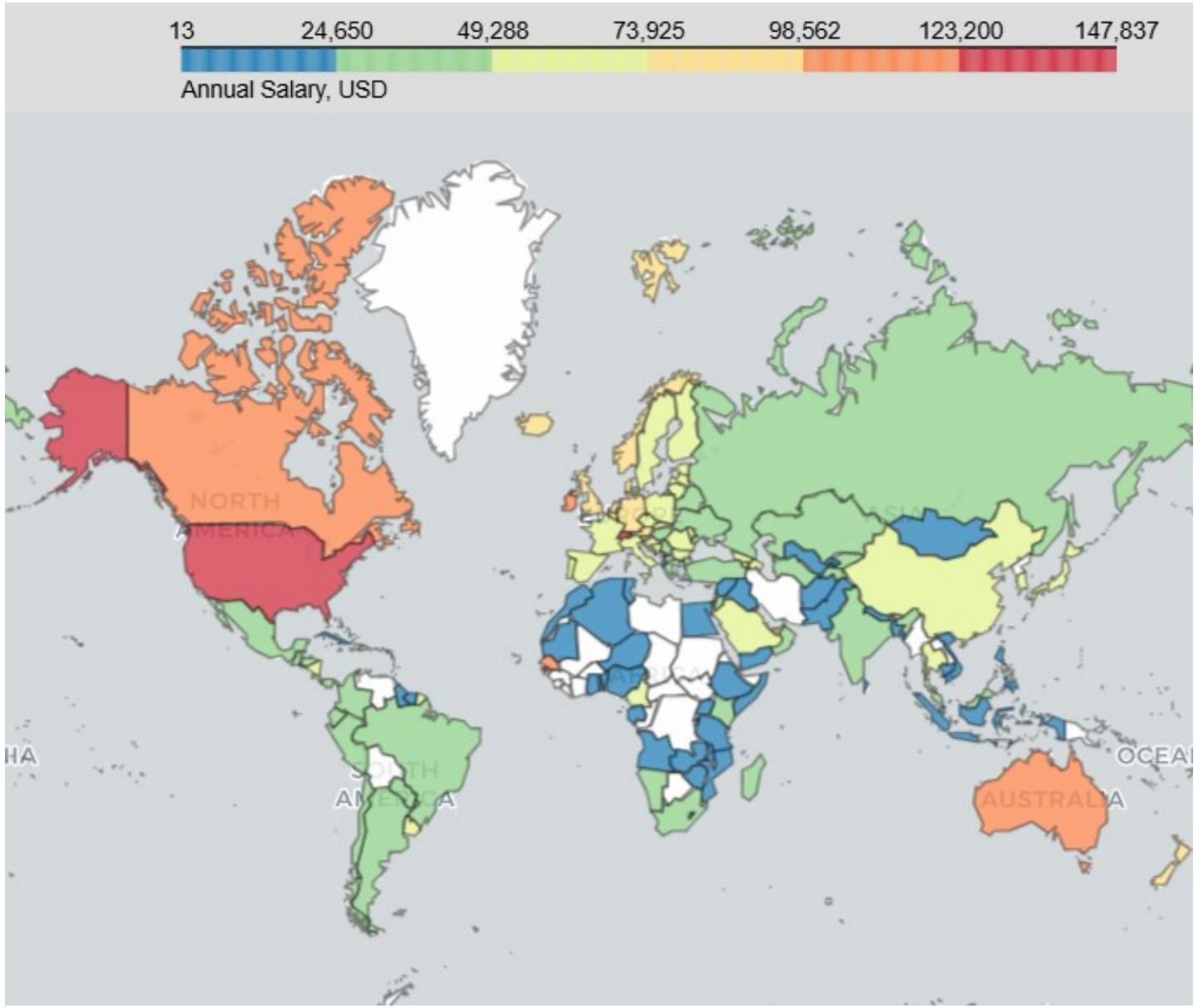
Imbalance in Data

- Not all countries have the same representation in data
- Most active part of community lives in the USA, Germany, UK, India and Canada
- Data from Cuba, Iran, North Korea, and Syria are inaccessible

Region	Respondents
Northern America	27.69%
Western Europe	17.81%
Northern Europe	13.53%
Eastern Europe	8.61%
Southern Europe	7.88%
Southern Asia	6.59%
South America	4.92%
Australia and New Zealand	3.13%
Western Asia	3.04%
South-eastern Asia	1.93%
Eastern Asia	1.34%
Central America	1.17%

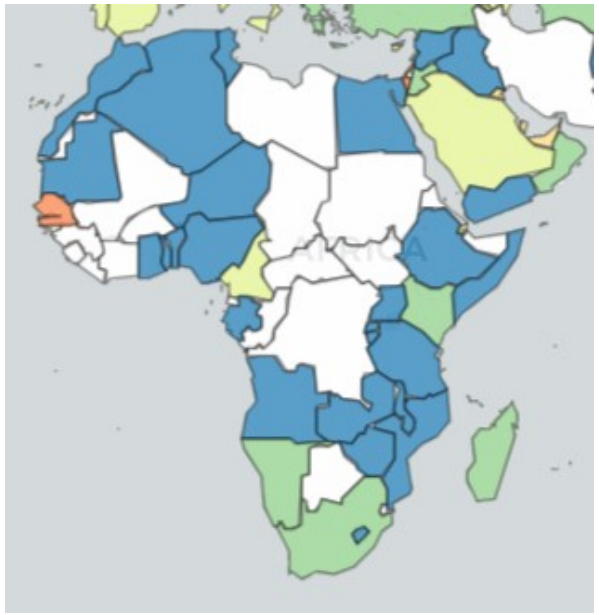
Salary by country and region

Region	Average Annual Salary
Northern America	$140604
Australia and New Zealand	$94708
Northern Europe	$83519
Western Europe	$78772
Western Asia	$69702
Eastern Asia	$58071
Southern Europe	$54929
Eastern Europe	$54568
Southern Africa	$47852
Central America	$47231
South America	$40974
South-eastern Asia	$37288
Caribbean	$37160
Middle Africa	$33598
Central Asia	$30549
Southern Asia	$23328
Western Africa	$21044
Eastern Africa	$20598
Melanesia	$19438
Micronesia	$14704
Northern Africa	$11579



Senegal: outlier

Western Africa	$21044
Eastern Africa	$20598
Melanesia	$19438
Micronesia	$14704
Northern Africa	$11579



EdLevel	YearsCodePro	LanguageHaveWorkedWith	DevType	Country	Region	Industry
Master	6	HTML/CSS;JavaScript;SQL;TypeScript	Developer, full-stack	Senegal	Western Africa	Manufacturing, Transportation, or Supply Chain

Highest Paying

	2023 Avg. Annual Salary
Marketing or sales professional	$118777
Engineering manager	$123657
Engineer, site reliability	$122281

Employment Roles: 34 different roles

	Non-graduated	Bachelor	Master	PhD
Marketing or sales professional	31.25%	37.50%	25.00%	6.25%
Engineering manager	16.96%	47.53%	32.18%	3.34%
Engineer, site reliability	26.23%	46.45%	25.68%	1.64%

Highest Paying Roles: Education

Top10 Programming languages

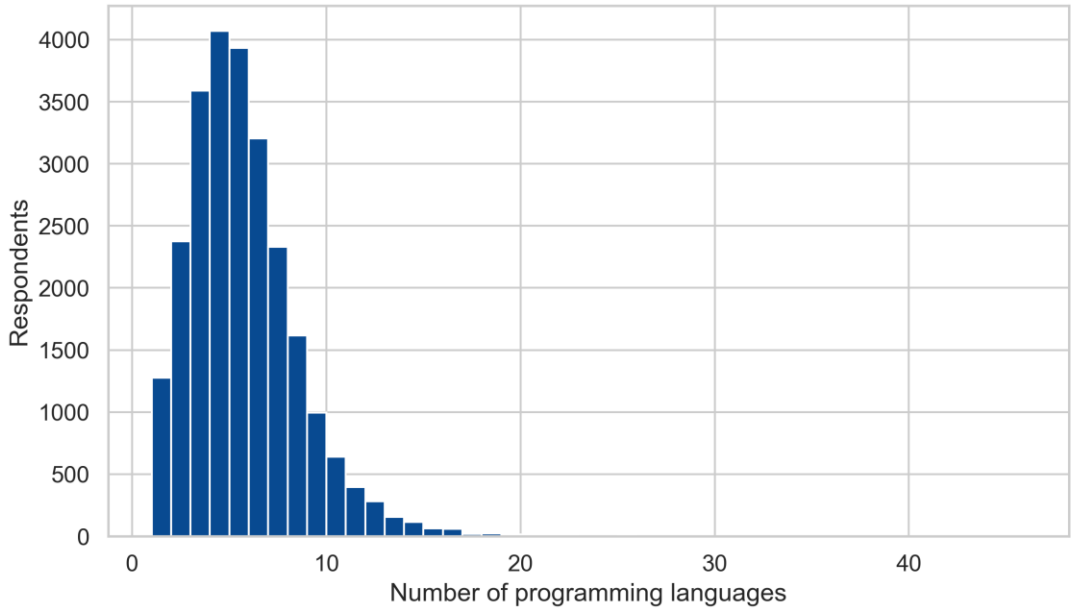
Python
SQL
Bash/Shell (all shells)
HTML/CSS
JavaScript
R
C++
Java
TypeScript
C#

	2023 Avg. Annual Salary
Data or business analyst	$69249
Data scientist or machine learning specialist	$89384

	Non-graduated	Bachelor	Master	PhD
Data or business analyst	17.22%	40.56%	37.78%	4.44%
Data scientist or machine learning specialist	2.33%	26.02%	52.43%	19.22%

Analytics employment roles

Most commonly-used programming language

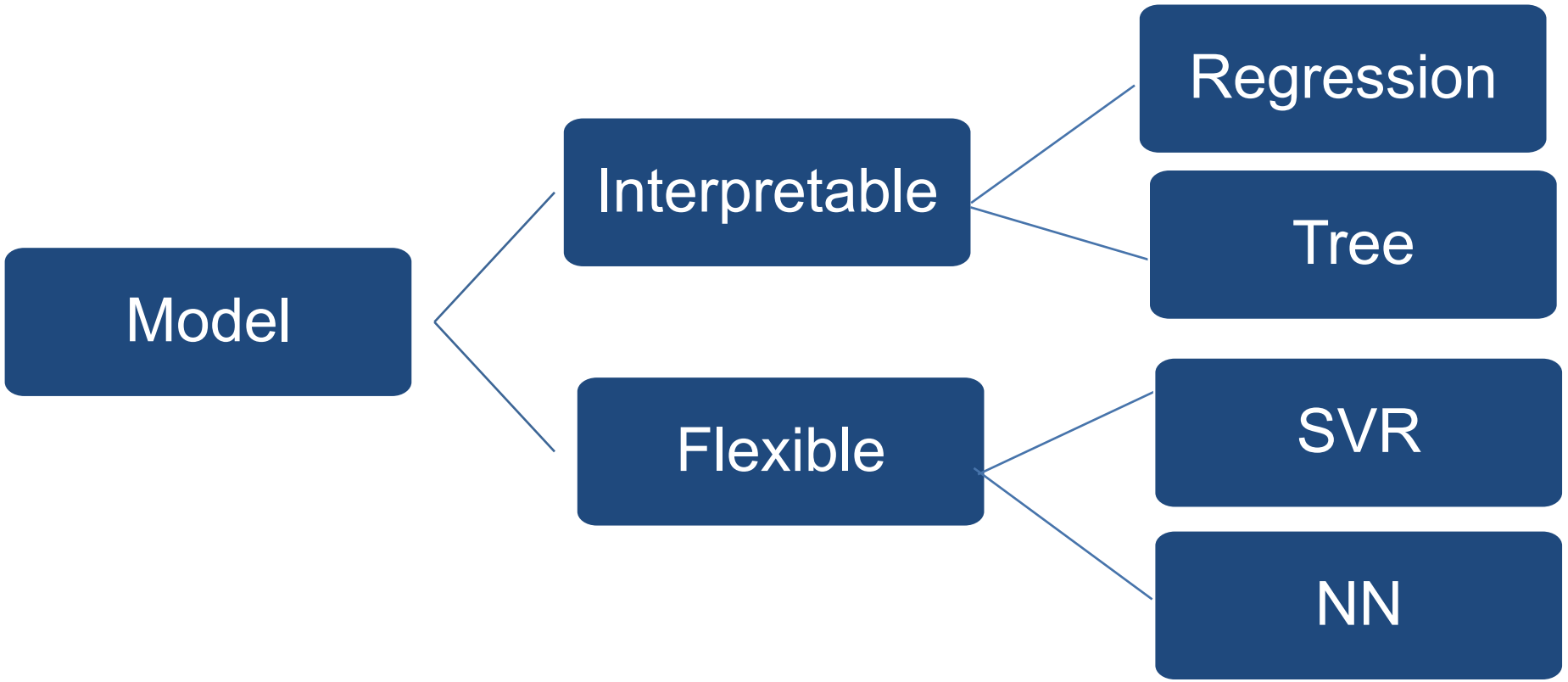


	Respondents
JavaScript	65%
SQL	53%
HTML/CSS	53%
TypeScript	46%
Python	45%
Bash/Shell (all shells)	36%
C#	29%
Java	28%
PHP	17%
C++	17%
PowerShell	15%
Go	15%

Analytics Project: Predicting Salary

Develop a machine-learning model to **predict annual salary** based on various factors such as:

- *Country*
- *Employment role*
- *Education level*
- *Years of professional coding*
- *Most used programming languages*



Selecting a model to predict salary

Linear Regression

- Outliers have a strong influence on the regression model
- Do not fit well complex patterns
- Simple
- Computationally efficient
- Multicollinearity between independent variables

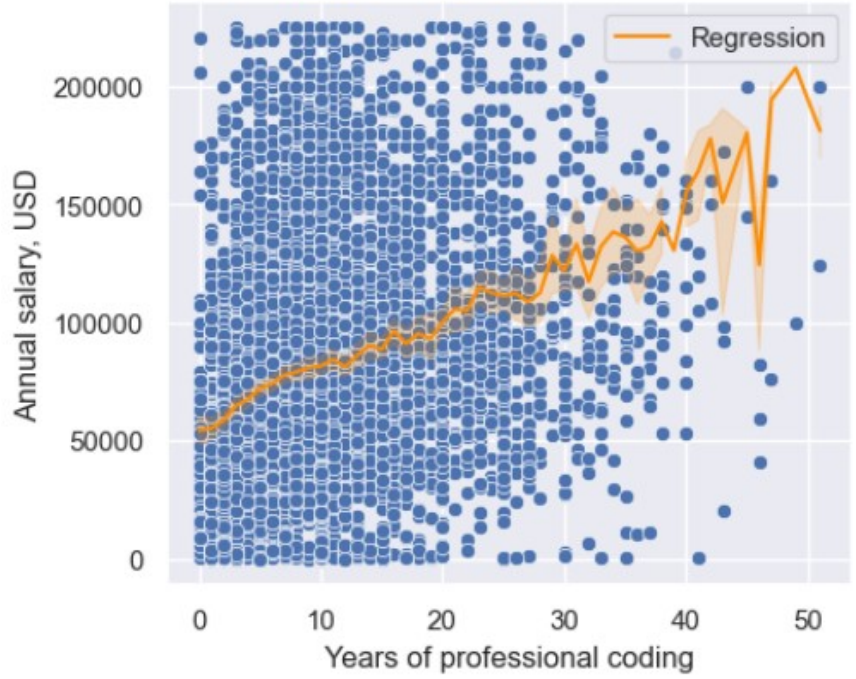
Regression Tree

- Robust to outliers in the input features
- Imbalanced classes and high cardinality make the model training process more challenging
- Outliers in target variable can affect results
- Flexibility and complexity
- Good for data with many categorical variables

Selecting a model to predict salary

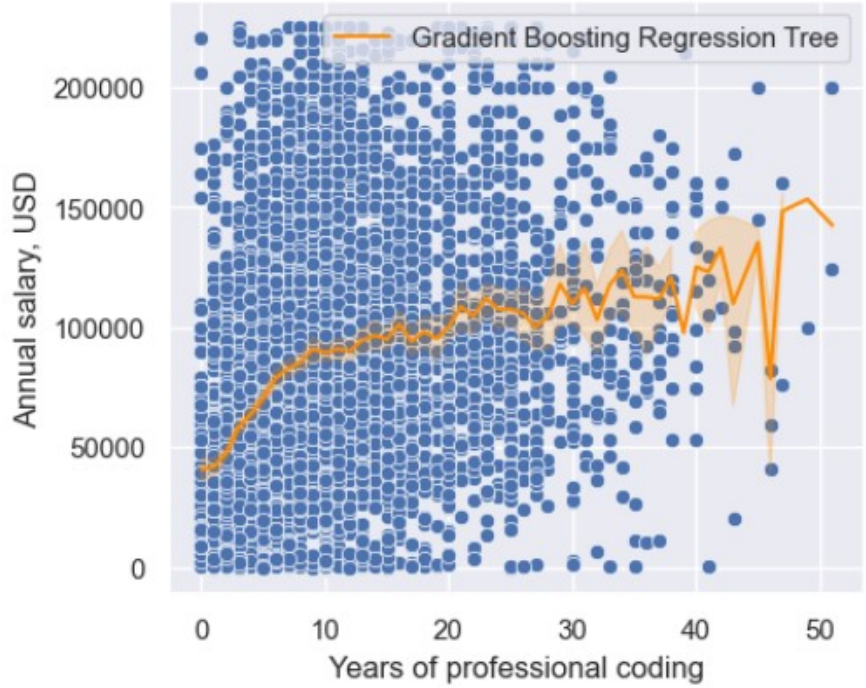
Linear Regression

Model performance on train data R^2 = 0.635
Model performance on test data R^2 = 0.630



Gradient Boosting Regression Tree with tuned hyperparameters

Model performance on train data R^2 = 0.689
Model performance on test data R^2 = 0.660



Comparing models

Salary prediction: data scientist vs data analyst

Data Scientist

$79255

90% Prediction Interval

[$52443; $93526]

Education: **Master**

Coding Experience: **5**

Country: **Germany**

Programming Languages: **Python, R, SQL**

Data Analyst

$61805

90% Prediction Interval

[$47423; $85432]

Analytics Project: Predicting Salary

Takeaways

- Prediction annual salary is based on location, experience and skills
- Choose an algorithm that balances complexity, flexibility, and predictive power
- Affordable computational time to fit a model
- Further steps: adding new features and improving model accuracy

THANK YOU

Lena Smotrova

www.linkedin.com/in/lena-smotrova/