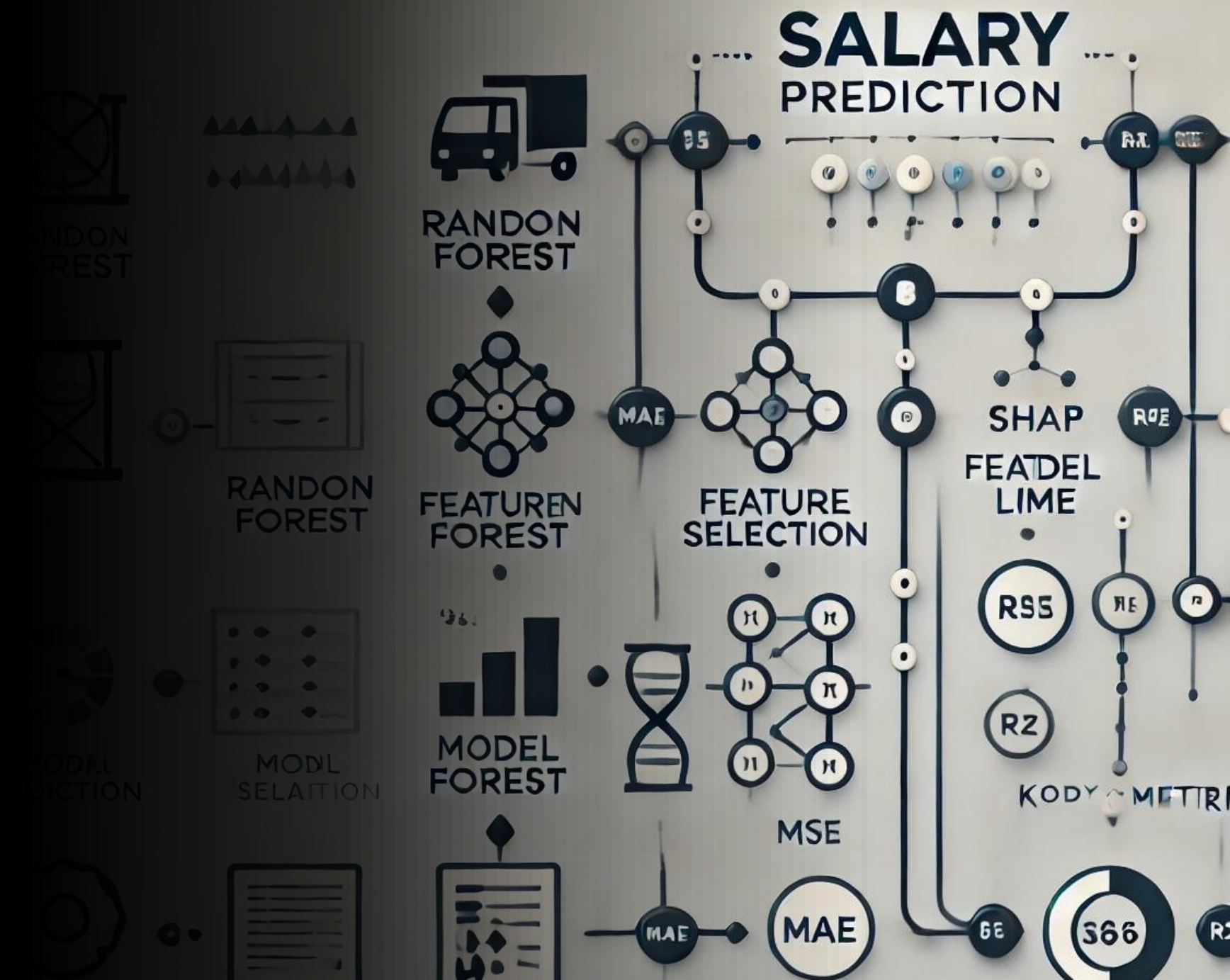
# PROJECT OF AI FOR CYBERSECURITY

Giacomo Maldarella




University of Pisa
AI4Cybersec 931II
09-01-2025

# Table of Content


Goal and specifics

Data Acquisition

Data Exploration

Preprocessing

Processing

Result Summary

Acknowledgments

# Goal and specifics

**Goal:** Develop a machine learning model to **predict salaries** in the cybersecurity industry based on various features, enhancing accuracy and identifying key factors influencing salary variations.

**Specifications:**

1. **Data Exploration**
   - Visualize numerical and categorical features
   - Analyze relationships between features and salary
   - Explore trends over time and regional disparities
2. **Data Preprocessing**
   - **Data cleaning**: remove duplicates, handle missing values
   - **Encoding**: OneHotEncoder and OrdinalEncoder
   - **Scaling**: Standard scaling for numerical features
   - **Outlier detection**: Apply log transformation to target
3. **Model Development and Validation**
   - Models used:
     - **Linear Regression**
     - **Random Forest**
     - **XGBoost**
4. **Feature selection using:**
   - **Recursive Feature Elimination (RFE)**
5. **Explainable Artificial Intelligence**
   - **SHAP** (SHapley Additive Explanations)
   - **LIME** (Local Interpretable Model-agnostic Explanations)

Hyperparameter tuning with **RandomSearchCV**

**Key Metric Goals:**

Evaluate models using **K-Fold Cross Validation**

Focus on improving:

- Mean Absolute Error (**MAE**)
- Root Mean Squared Error (**RMSE**)
- R-Squared (**$R^2$**)

# Data Acquisition

## Dataset

salaries

| | work_year | experience_level | employment_type | job_title | salary | salary_currency | salary_in_usd | employee_residence | remote_ratio | company_location | company_size |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2024 | MI | FT | Security Consultant | 211000 | USD | 211000 | US | 0 | US | M |
| 3 | 2024 | MI | FT | Security Consultant | 142000 | USD | 142000 | US | 0 | US | M |
| 4 | 2024 | MI | FT | Security Consultant | 64417 | GBP | 80521 | GB | 0 | GB | M |
| 5 | 2024 | MI | FT | Security Consultant | 52584 | GBP | 65730 | GB | 0 | GB | M |
| 6 | 2024 | MI | FT | Consultant | 188400 | USD | 188400 | US | 0 | US | M |
| 7 | 2024 | MI | FT | Consultant | 125600 | USD | 125600 | US | 0 | US | M |
| 8 | 2024 | MI | FT | Manager | 246400 | USD | 246400 | US | 0 | US | M |
| 9 | 2024 | MI | FT | Manager | 117300 | USD | 117300 | US | 0 | US | M |
| 10 | 2024 | MI | FT | Security Engineer | 200200 | USD | 200200 | US | 0 | US | M |
| ... | | | | | | | | | | | |
| 22600 | 2020 | SE | FT | Information Security Officer | 35000 | USD | 35000 | AR | 100 | AR | L |
| 22601 | 2021 | EN | FT | Application Security Engineer | 65000 | USD | 65000 | US | 100 | US | L |
| 22602 | 2020 | SE | FT | Information Security Specialist | 170000 | USD | 170000 | US | 100 | US | L |
| 22603 | 2021 | SE | FT | Application Security Engineer | 135000 | USD | 135000 | US | 100 | US | L |
| 22604 | 2021 | EN | FT | Cyber Security Analyst | 100000 | USD | 100000 | US | 50 | US | M |
| 22605 | 2020 | MI | FT | Ethical Hacker | 356000 | GBP | 456621 | GB | 100 | GB | L |
| 22606 | 2020 | MI | FT | Cyber Security Analyst | 140000 | AUD | 96422 | AU | 50 | AU | M |
| 22607 | 2021 | SE | FT | Information Security Manager | 60000 | GBP | 82528 | GB | 50 | GB | L |
| 22608 | 2021 | SE | FT | Penetration Testing Engineer | 126000 | USD | 126000 | US | 100 | US | L |
| 22609 | 2021 | MI | FT | Information Security Analyst | 42000 | GBP | 57769 | GB | 100 | GB | L |
| 22610 | 2021 | MI | FT | Threat Intelligence Analyst | 66310 | USD | 66310 | US | 0 | US | L |

**Platform:** Kaggle

## Reason for Choosing This Dataset:

The dataset provides insights into the evolving landscape of cybersecurity roles, highlighting salary disparities based on experience, location, and job titles.

## Dataset Description:

The dataset contains over **20,000 records** from various countries, with a strong prevalence of data from the United States.

It spans from **2020 to 2024**, making it highly recent and relevant to current industry trends.

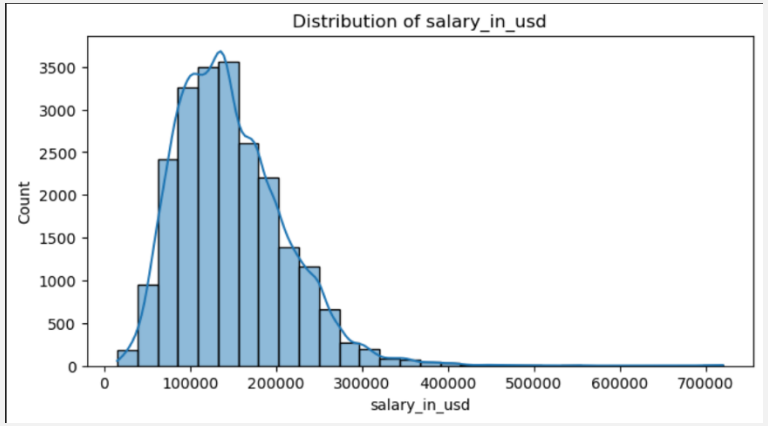
The data includes both **continuous** and **categorical variables**.

# Data Exploration

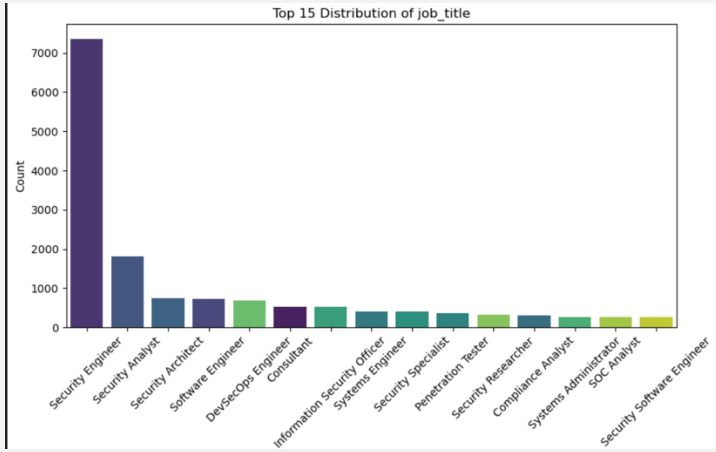
**Objective:**

Understand the dataset through visual and statistical analysis to identify patterns, correlations, and key insights that may influence salary predictions.

**1. Visualizing Numerical Features**




**2. Visualizing Categorical Features**



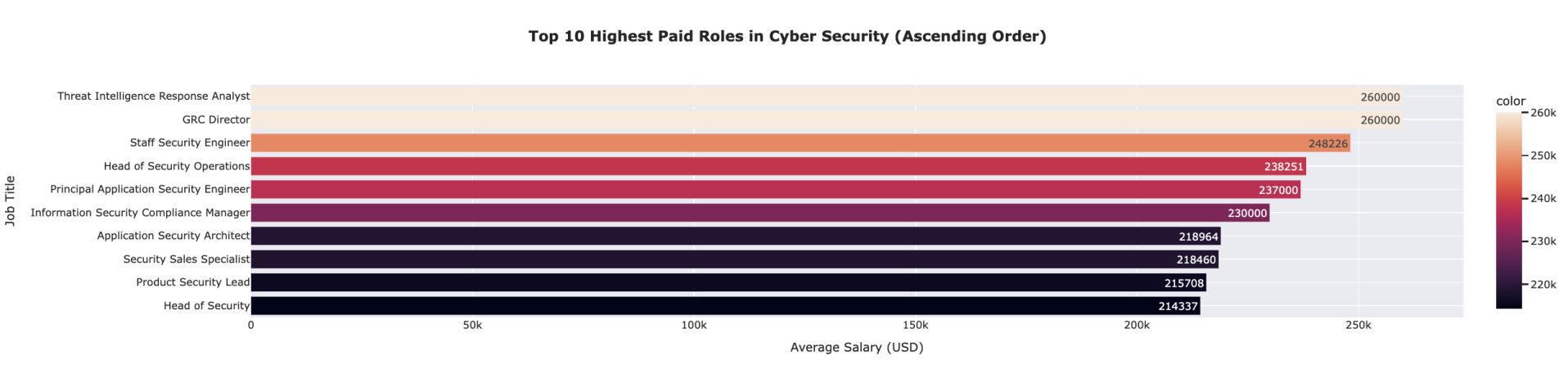

**3. Relationship Between Categorical Features and Salary**

**4. Relationship Between Numerical Features and Salary**

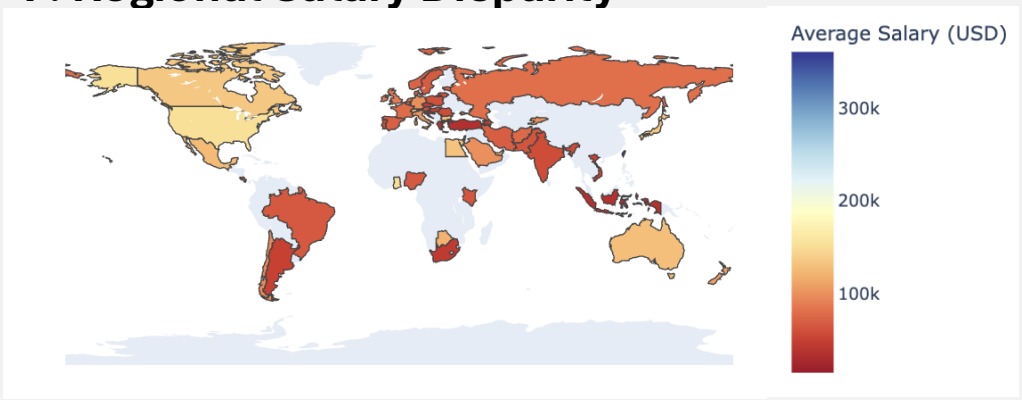
**5. Salary Trends Over Time**

# Data Exploration

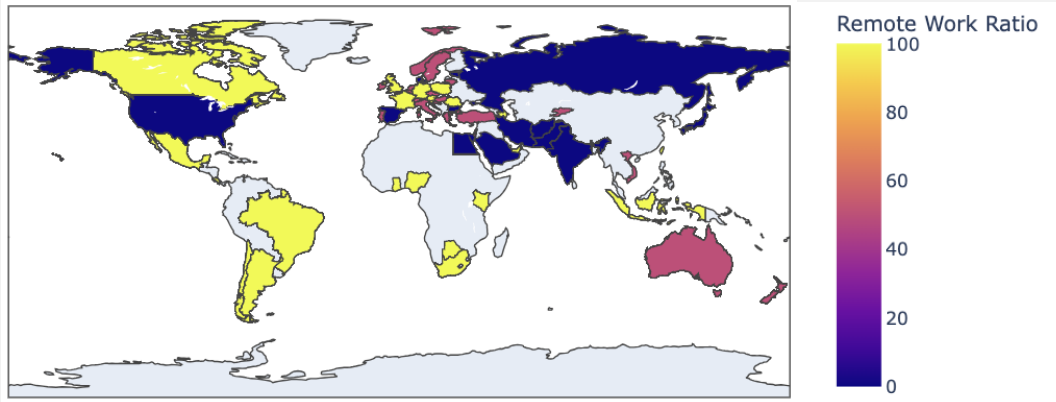
## 6. Average Salary for Different Cybersecurity Roles




## 7. Regional Salary Disparity




## 8. Remote work and salary



# Data Preprocessing – Data Cleaning

**Objective:**

Ensure data quality by addressing duplicates, missing values, and high-cardinality features to improve model performance.

## Duplicates Removal

```
Checking for Duplicates...
Number of duplicate rows: 9333
Duplicate rows removed.
```

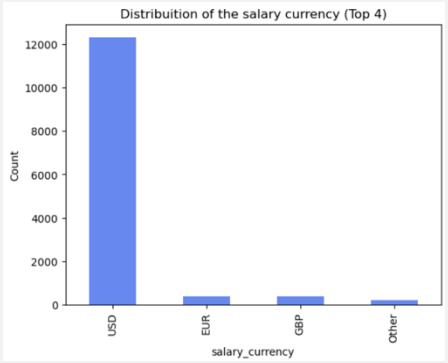
Identified and removed duplicate records to ensure unique observations.

## Handling Missing Values:

```
Final Check for Missing Values:
work_year              0
experience_level       0
employment_type        0
job_title              0
salary                 0
salary_currency        0
salary_in_usd          0
employee_residence     0
remote_ratio           0
company_location       0
company_size           0
dtype: int64
```

Missing values were imputed or dropped depending on their relevance to the analysis.

## Reducing High Cardinality in Categorical Features:




1. For categorical columns **with numerous unique values** (e.g., employee_residence), only the top 4 most frequent categories were retained.
2. All remaining categories were grouped under "Other" to simplify the dataset and prevent overfitting.

# Data Preprocessing - Encoding

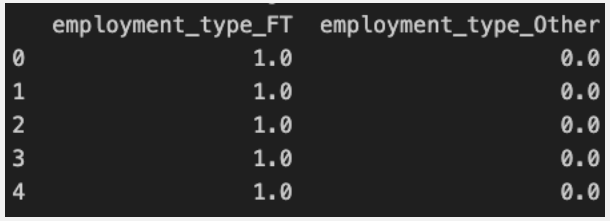
**Objective:**
Convert categorical features into a numerical format suitable for machine learning models.

**1.OneHotEncoder** – Applied to features with a small number of unique values.

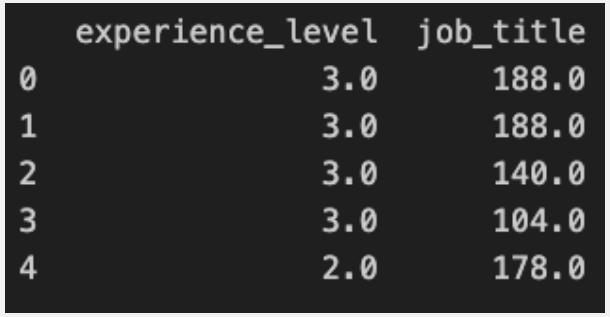
1. Each unique category is represented as a binary column (0 or 1).
2. Prevents losing valuable information from low-cardinality features without introducing bias.

**2.OrdinalEncoder** – Used for features with a logical order.

1. This method assigns an integer to each category, preserving its rank or significance.
2. helps manage high-cardinality features by avoiding an explosion of dimensions, keeping the dataset compact and manageable.



| | employment_type_FT | employment_type_Other |
|---|---|---|
| 0 | 1.0 | 0.0 |
| 1 | 1.0 | 0.0 |
| 2 | 1.0 | 0.0 |
| 3 | 1.0 | 0.0 |
| 4 | 1.0 | 0.0 |



| | experience_level | job_title |
|---|---|---|
| 0 | 3.0 | 188.0 |
| 1 | 3.0 | 188.0 |
| 2 | 3.0 | 140.0 |
| 3 | 3.0 | 104.0 |
| 4 | 2.0 | 178.0 |

# Data Preprocessing – Outlier detection and scaling

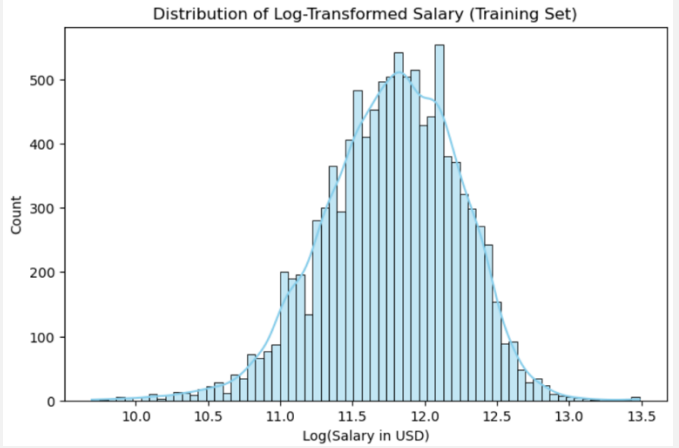
**Objective:**
Normalize the data to ensure models are not biased by extreme values and to bring numerical features to the same scale.

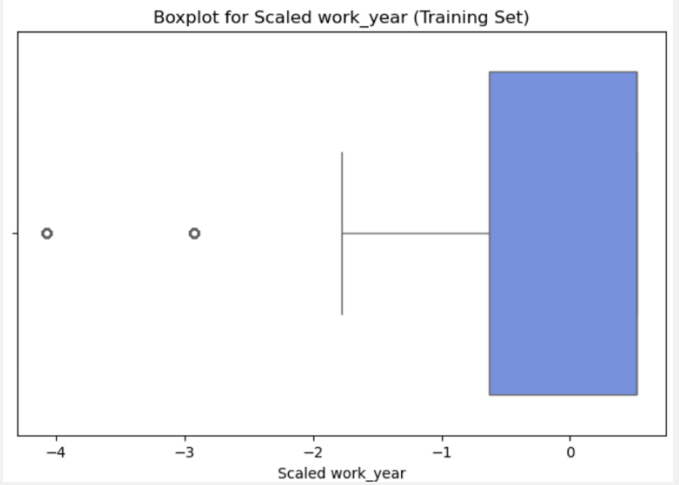
**1.Outlier Detection (Log Transformation):**

1. A log transformation was applied **only to the target variable** (salary_in_usd).
2. This helps reduce the skewness caused by extreme salary values, stabilizing model predictions.

**2.Scaling Numerical Features:**

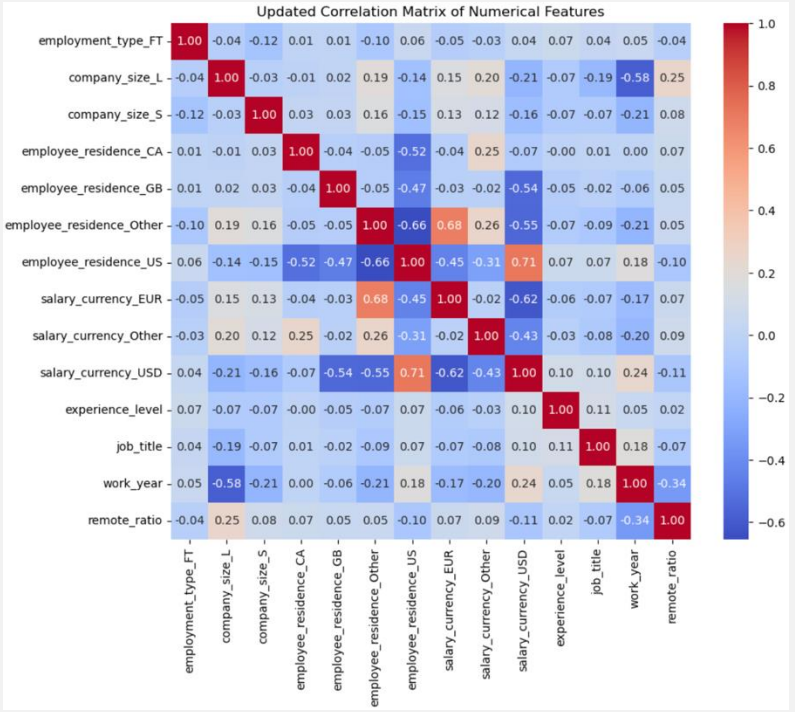
1. StandardScaler was applied to **work_year** and **remote_ratio**.
2. These features have different ranges, and scaling ensures the model treats them equally during training.

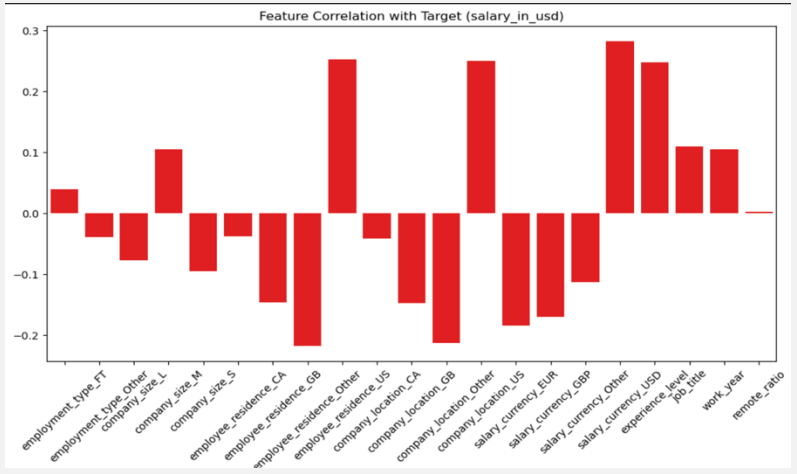


# Data Preprocessing – Pearson Correlation

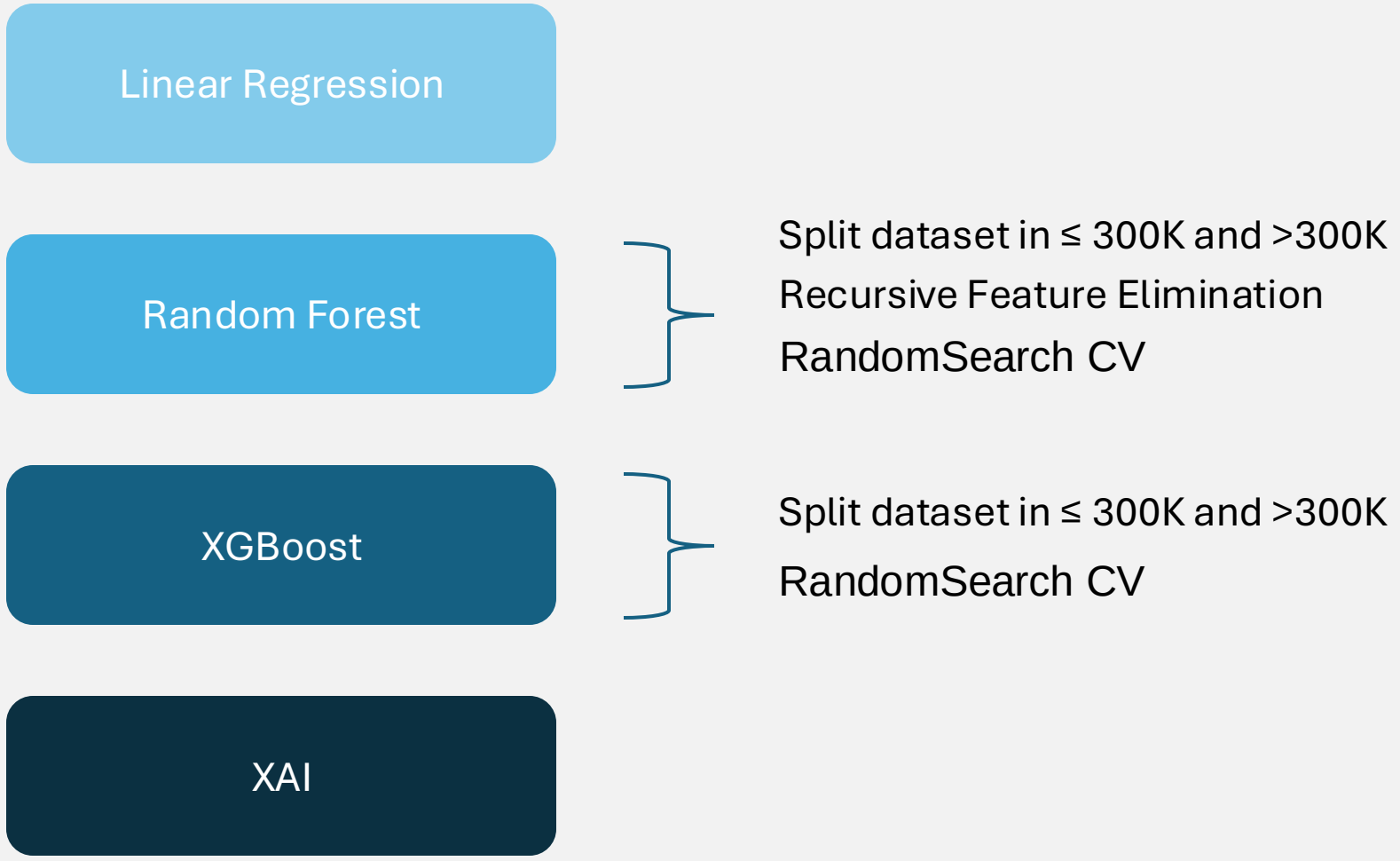
- **Correlation Matrix for Numerical Features:**
  - A heatmap was generated to visualize correlations between numerical features.
  - This helps identify relationships between independent variables.
- **Correlation with Target (salary_in_usd):**
  - A bar plot shows how each feature correlates with the target.
  - Positive correlations indicate features that increase with salary, while negative values suggest the opposite.
- **Threshold for High Correlation:**
  - Pairs of features with a correlation higher than **0.85** were flagged.
- **Results:**
- Features such as **salary_currency_USD** and **employee_residence_US** show the strongest positive correlation with salary_in_usd.
- No features had extreme correlations above the threshold, indicating no immediate need to drop significant columns.
- This analysis ensures model stability and avoids overfitting by reducing redundant data.



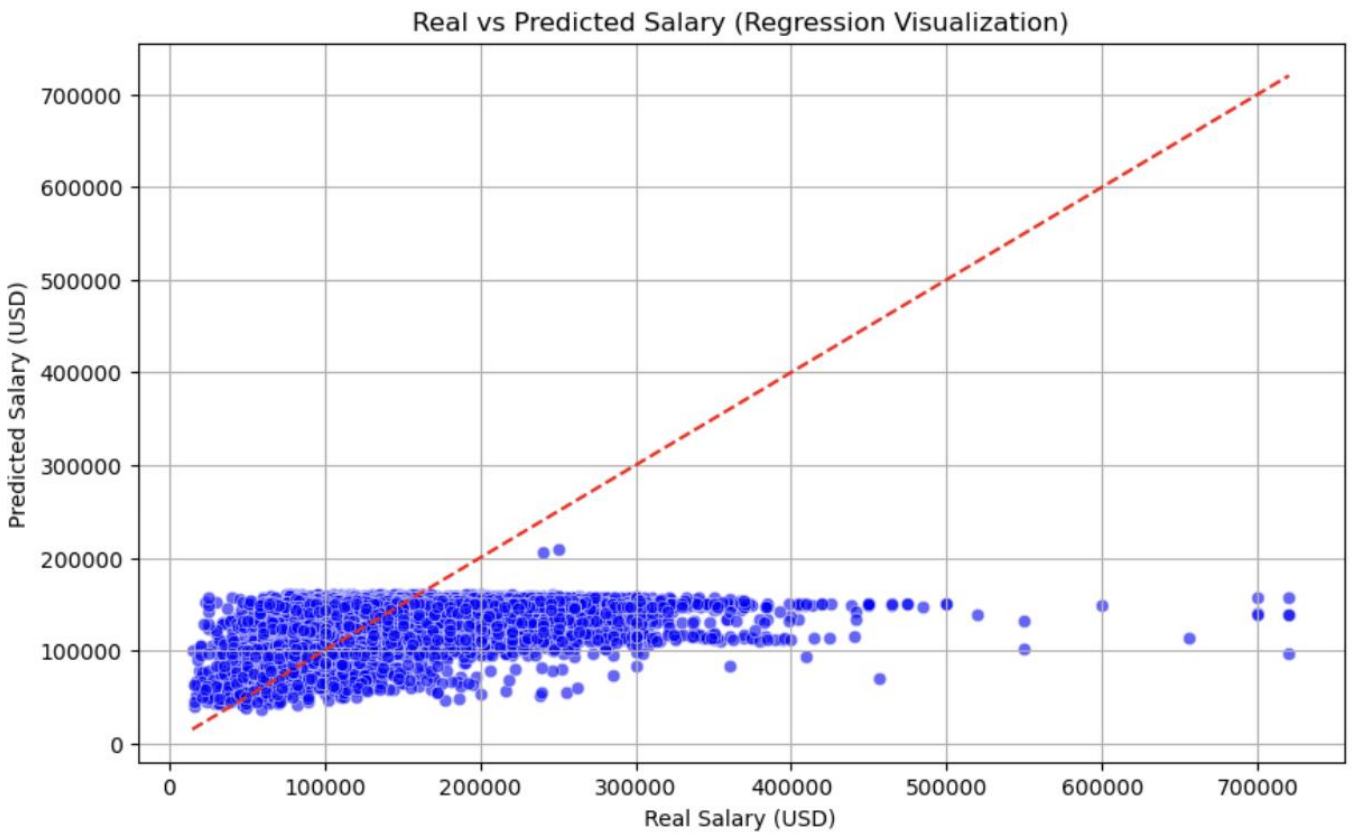



# Data Processing

**Overall processing**



# Data Processing – Linear Regression




**Results Interpretation:**

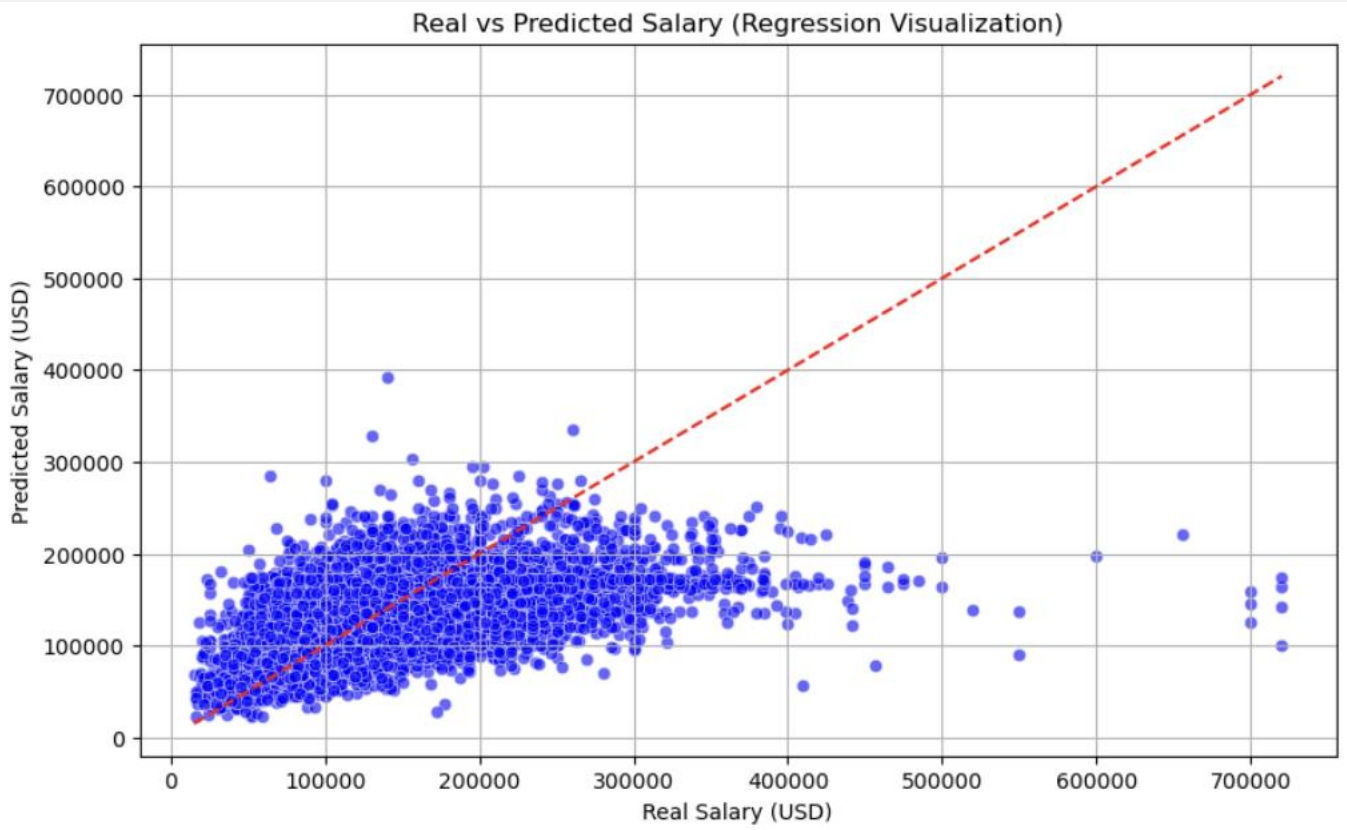
•**Visual Analysis (Scatter Plot)**:

- The red dashed line represents the **ideal prediction (y = x)**.
- Most predictions cluster below the line, indicating the model tends to **underestimate higher salaries**.
- There are a few extreme outliers (salaries > 300k) poorly predicted by the model.

•**Performance Metrics (Bottom Panel):**

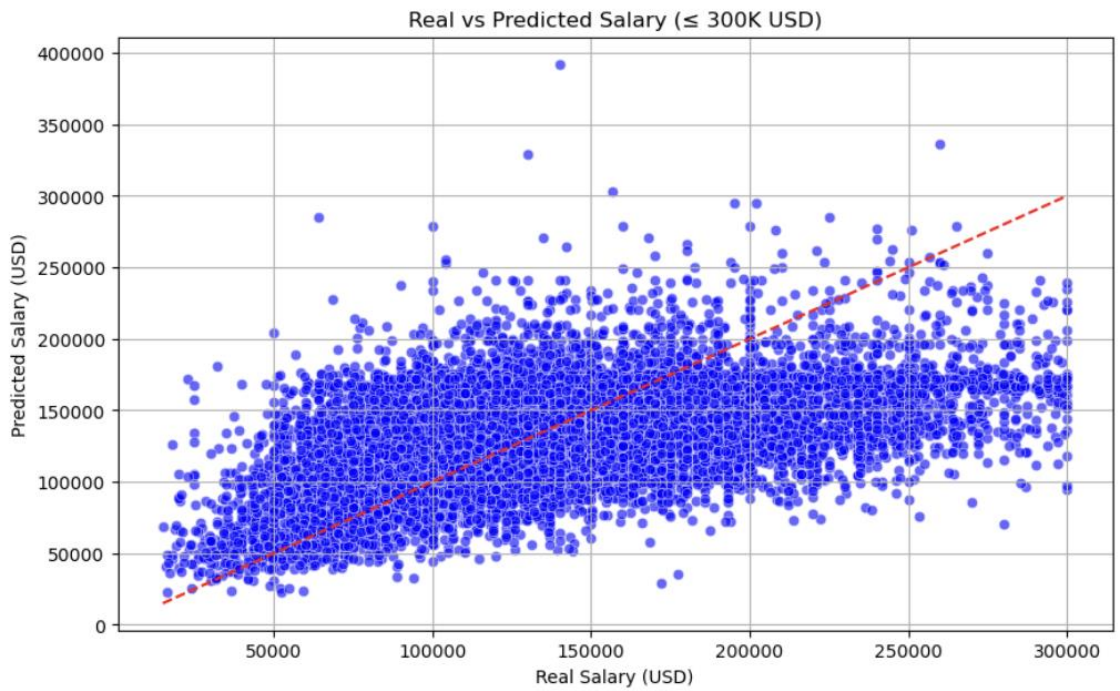
- **MAE**: 44,758 USD (average error in salary prediction).
- **RMSE**: 61,631 USD (penalizes larger errors).
- **$R^2$**: 0.11 (the model explains only **11% of variance**, indicating low performance).

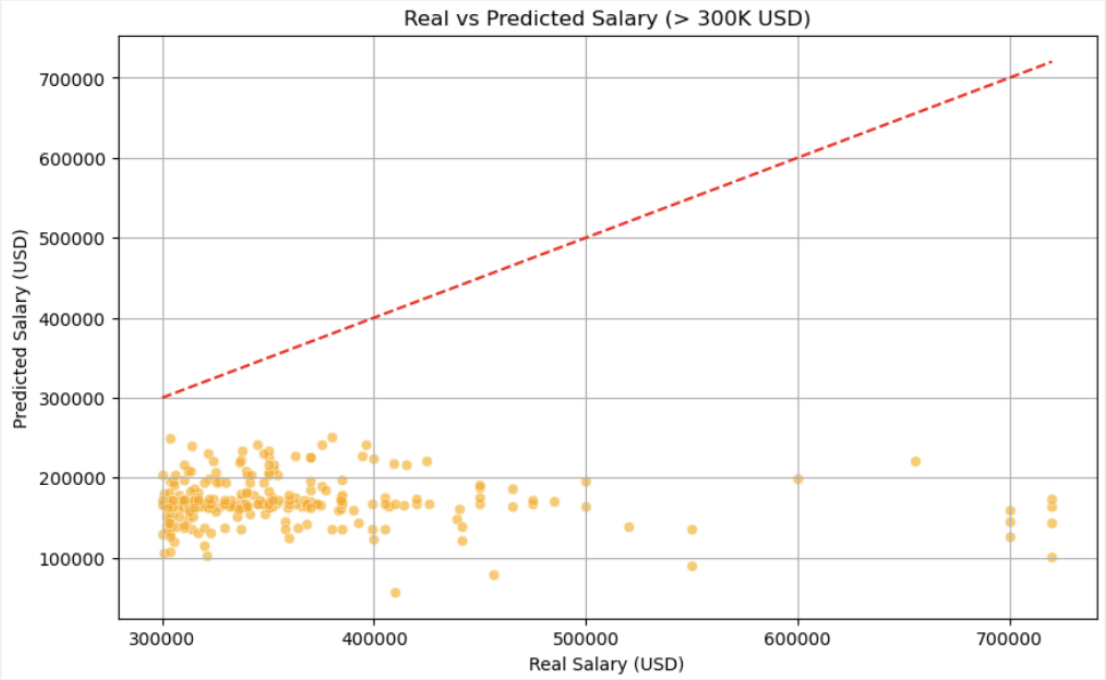
# Data Processing – Random Forest




1. **General Trend:**
    1. The model captures some degree of salary variation, but not with high precision.
2. **Deviations and Outliers:**
    1. For salaries above approximately $300,000, predictions tend to flatten, suggesting **underestimation** of high salaries.
    2. Outliers are present, particularly at higher salary ranges, where the model struggles to make accurate predictions.
3. **Evaluation Metrics:**
    1. **Mean Absolute Error (MAE):** $43,018
    2. **Root Mean Squared Error (RMSE):** $57,973
    3. **$R^2$ (R-Squared):** 0.21 – Indicates that only **21% of the variance** in salaries is explained by the model. This is relatively low, showing room for improvement.

# Data Processing – Random Forest  split value




**Results:**

- **Salaries ≤ $300K:**
  - Model shows reasonable predictive capability with an $R^2$ of **0.22**.
  - Errors are distributed relatively consistently, but some underestimation is observed at higher salary levels.
- **Salaries > $300K:**
  - Model struggles significantly for high salaries, with an $R^2$ of **-6.04**.
  - This suggests the need for more training data or different modeling approaches for extreme salaries.

# Data Processing – Feature Selection

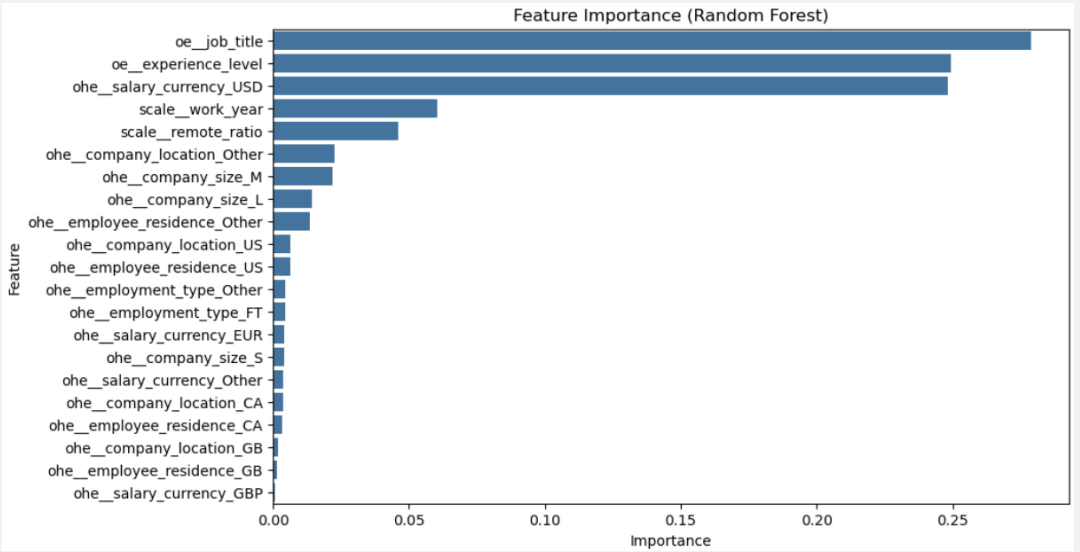
### 1. Random Forest Feature Importance

**•How it Works**:

- Random Forest provides a built-in feature importance mechanism by calculating how much each **feature decreases the impurity (Gini or variance).**

**•Results**:

- The most important features were:
  - **oe__job_title - ohe__salary_currency_USD- oe__experience_level**
- Other factors, such as **company size** and **location**, had lower importance.




### 2. Permutation Importance

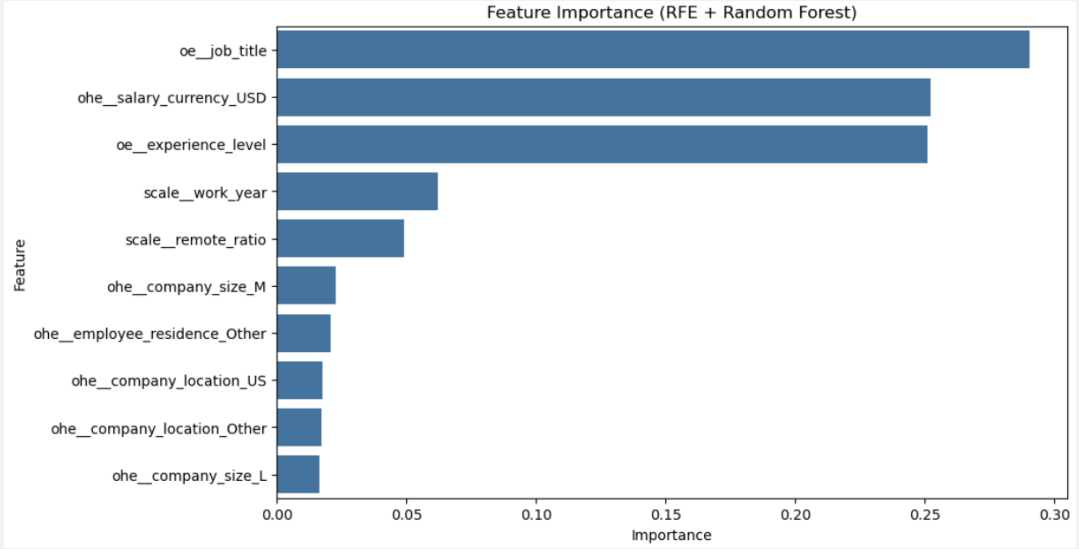
### 3. Recursive Feature Elimination (RFE)

**•How it Works**:

- RFE recursively **removes** the **least important features** by training the model **iteratively**. It continues until the desired number of features is selected.

**•Results**:

- The top 10 features selected through RFE align with the results of Feature Importance and Permutation Importance.
- This method reinforced the significance of **job title**, **salary currency**, and **experience level**.



# Data Processing – RandomSearch CV

**Hyperparameters Explained:**

**1.n_estimators** – Number of trees in the forest.

1. Values tested: **[100, 200, 300, 400]**
2. **Best Value**: **300**

**2.max_depth** – Maximum depth of the individual decision trees. (Prevent overfitting)

1. Values tested: **[10, 20, 30, 50, None]**
2. **Best Value**: **30**

**3.min_samples_split** – Minimum number of samples required to split an internal node.

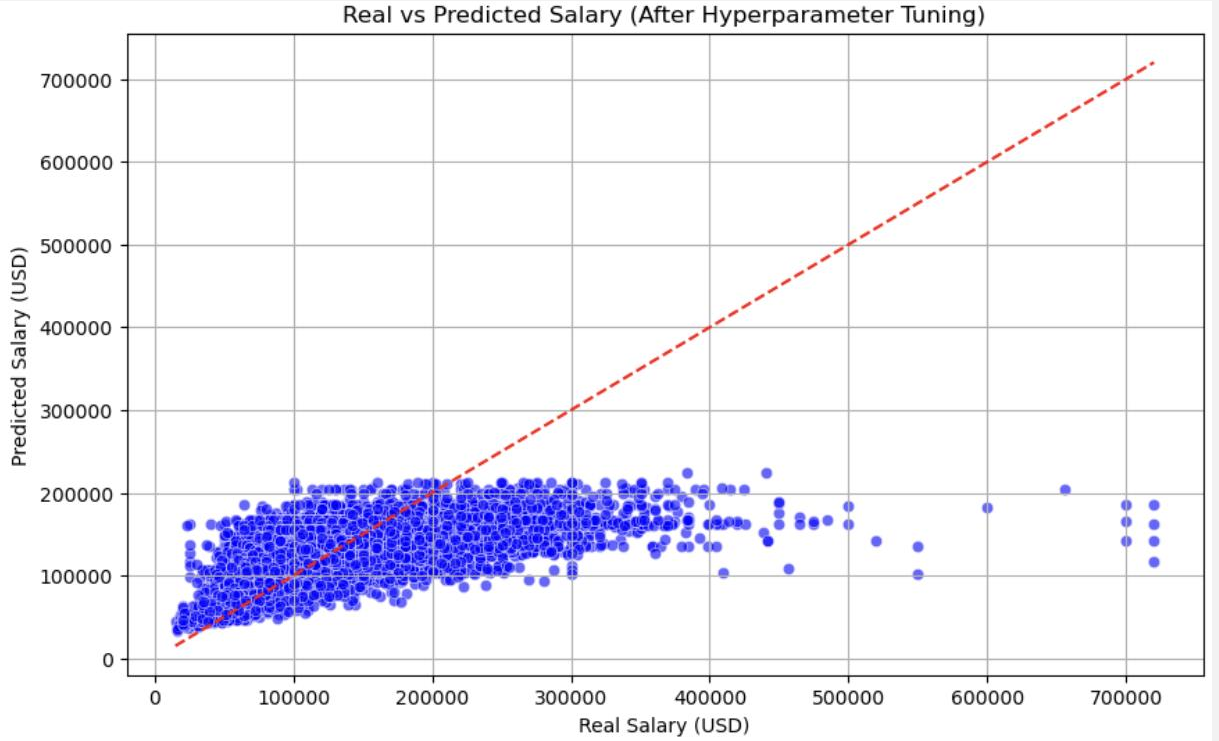
1. Values tested: **[2, 5, 10]**
2. **Best Value**: **2** (aggressive splits to prevent underfitting)

**4.min_samples_leaf** – Minimum number of samples required to be in a leaf node.

1. Values tested: **[1, 2, 4]**
2. **Best Value**: **2**

**5.max_features** – Number of features to consider for the best split.

1. Values tested: **[sqrt, log2]**
2. **Best Value**: **sqrt** (uses the square root of the total features, balancing performance and speed)

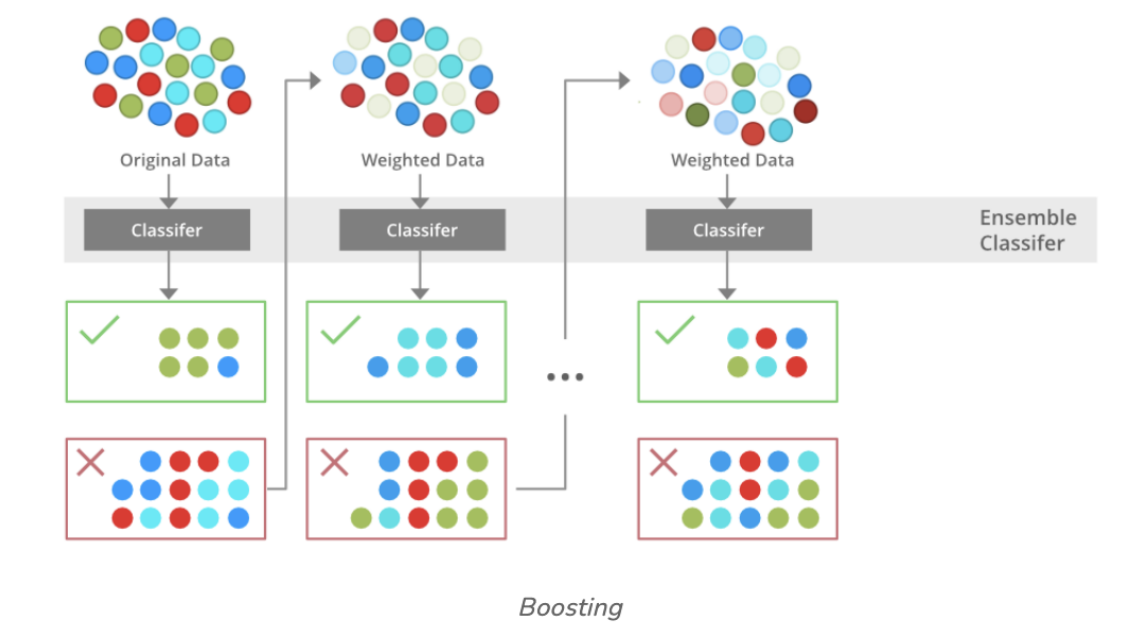



**Results After Tuning:**

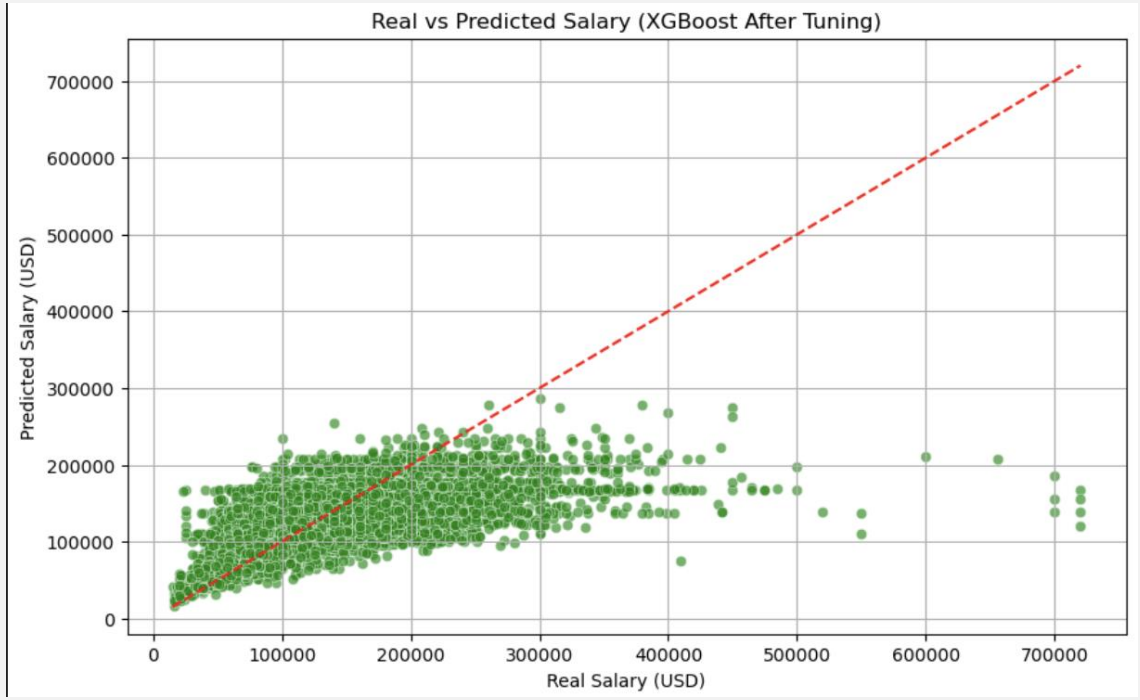
- **MAE**: **39,064** USD
- **RMSE**: **54,749** USD
- **$R^2$**: **0.2996**

# Data Processing – XGBoost RandomSearch CV

[1] Reference

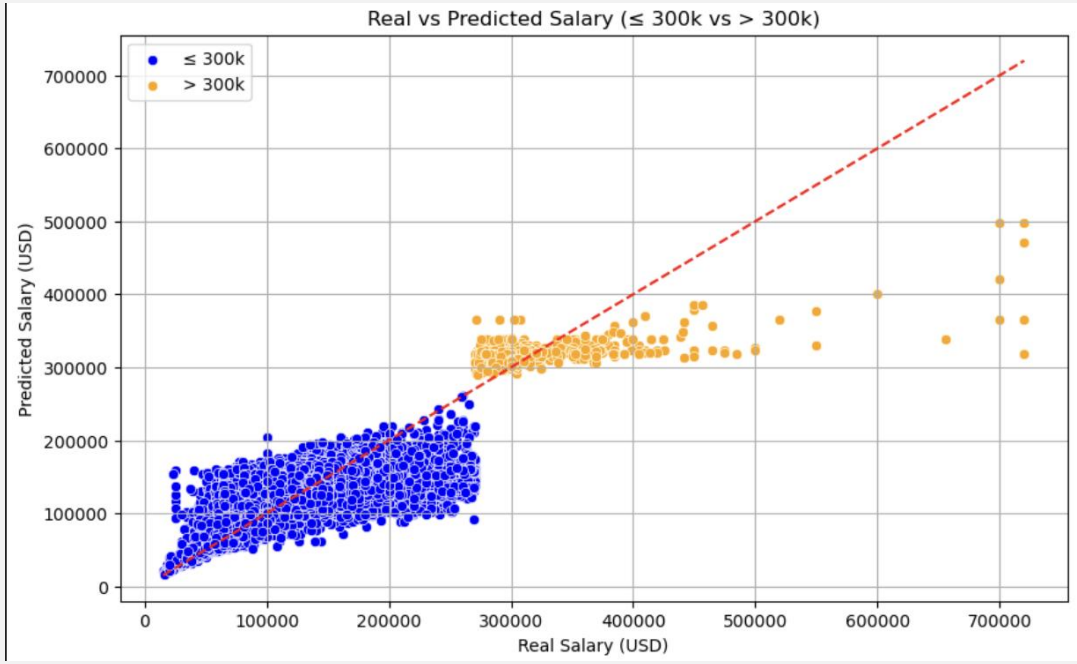



**1.First Classifier**: Trained on the original data. Misclassified samples are assigned **higher weights**.
**2.Second Classifier**: Focuses on **correcting the errors** from the first classifier. Misclassified points continue to receive more weight.
**3.Iterative Process**: This continues until a **strong ensemble classifier is created** by aggregating all weak learners.
**4.Final Model**: The ensemble classifier makes the **final decision**, **combining** the outputs of all classifiers.




**Results After Tuning:**
- **MAE**: **39,551** USD
- **RMSE**: **54,630** USD
- **$R^2$**: **0.3026**

# Data Processing – XGBoost split value




The model was divided into two segments:

**1.Salary ≤ 300k**:

1. XGBRegressor **without** hyperparameter tuning.
2. Faster training and evaluation.

**2.Salary > 300k**:

1. XGBRegressor **with** hyperparameter tuning using RandomizedSearchCV.
2. Aimed at addressing the higher variance in this segment.

•**Lower Salary Range (≤ 300k)**:

- **MAE**: 34,043 USD
- **RMSE**: 43,651 USD
- **$R^2$**: 0.35
- The model performs better for this segment, achieving higher $R^2$ and lower errors.

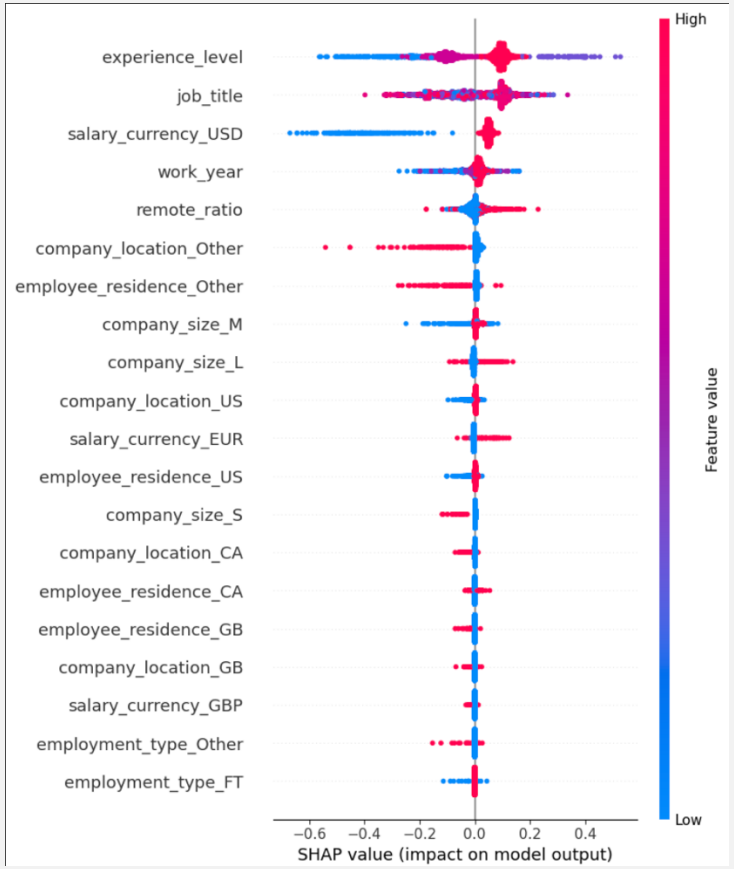
•**Higher Salary Range (> 300k)**:

- **MAE**: 34,931 USD
- **RMSE**: 54,747 USD
- **$R^2$**: 0.41
- Predictions for higher salaries show slightly higher errors and lower $R^2$, indicating challenges in modeling high-variance data.

# Data Processing – XAI

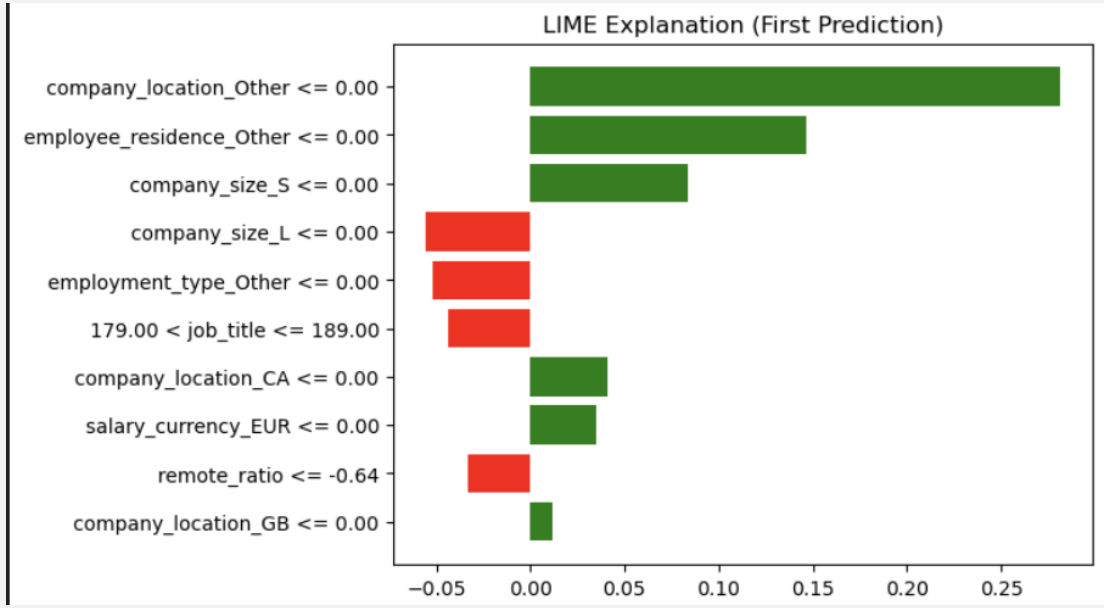
•**SHAP:**

- **Experience Level, Job Title, and Currency (USD)** are the most influential features and explainable of the model.
- Higher **remote ratio** negatively impacts predictions.




•**LIME:**

- Provides feature contributions for **individual predictions**.
- Local interpretability shows **job title** and **currency** as critical for salary predictions.



# Result Summary

| | MAE | RMSE | $R^2$ |
|---|---|---|---|
| Linear Regression | 44,758 | 61,631 | 0.11 |
| Random Forest | 43,018 | 57,973 | 0.21 |
| RFE | 43,050 | 57.005 | 0.21 |
| RF with RandomSearch CV | 39,064 | 54,749 | 0.29 |
| XGBoost with RandomSearch CV | **39,551** | **54,630** | **0.3026** |

| XGBoost | split | MAE | RMSE | $R^2$ |
|---|---|---|---|---|
| without | ≤300K | 34,043 | 43,651 | 0.35 |
| with | >300K | 34,931 | 54,747 | 0.41 |

- The results show that **XGBoost with RandomSearchCV** outperforms other models with the **lowest MAE (39,551)** and **RMSE (54,630)**, achieving the **highest $R^2$ (0.3026)**.

# Result Summary – T-test

**Approach:**

1.Conducted 5-fold cross-validation for each model.
2.Collected $R^2$ scores for each fold as the performance metric.
3.Applied **paired t-test** to compare the mean $R^2$ scores between models.

| Comparison | P-value | Significant |
|---|---|---|
| Linear Regression vs Random Forest | 0.0001 | Yes |
| Linear Regression vs XGBoost | 0.0002 | Yes |
| 0.0001 | 0.0024 | Yes |

**Conclusion:**

•**XGBoost** is the top-performing model, with statistically significant improvements over both **Random Forest** and **Linear Regression**.
•**Random Forest** also significantly outperforms **Linear Regression**, making it a strong alternative to XGBoost.

# Model Performance Comparison with Baseline and Next step

```
RMSE score for train 47,002 USD/year, and for test 52,431 USD/year
CPU times: user 2.88 s, sys: 460 ms, total: 3.34 s
Wall time: 1.4 s
```

```
RMSE baseline score for train 64,600 USD/year, and for test 65,261 USD/year
```

[2] Reference

**Conclusion:**

•While the RMSE values may seem high, they are **reasonable** given the nature of the dataset and the task's complexity.

•The model is **more effective** than the baseline, demonstrating a better generalization to unseen data.

•**Future improvements could involve:**

- **Collecting additional features** or external datasets.
- Implementing **advanced feature engineering** or domain-specific knowledge to improve model explainability.

# Thank you for the attention!

- Reference:

[1]: https://www.geeksforgeeks.org/xgboost/

[2]:https://www.kaggle.com/code/dima806/information-security-salary-autoviz-catboost-shap