JOB SALARY PREDICTION

TEXT ANALYTICS

Presented by: Hannah Wang JaeYoon Lee

PROBLEM STATEMENT

Current Challenge

Job seekers face difficulties in negotiating salaries
Employers struggle to benchmark compensation
Job platforms miss opportunities for enhanced recommendations

Solution

Develop a machine learning-based salary prediction model by leveraging NLP techniques such as bag-of-words, TF-IDF, and BERT

DATA PREPROCESSING

01

- Data Source: Linkedin Job Posting from Kaggle
- Columns: Title, Description, Company, Industry, Work Type, Experience Level, Salary
- Data Size: 27900

02

- Salary Range Adjustment
- Minimum Wage at U.S.: 7.5/hr -> 15000/year
- Max: Q3 + IQR

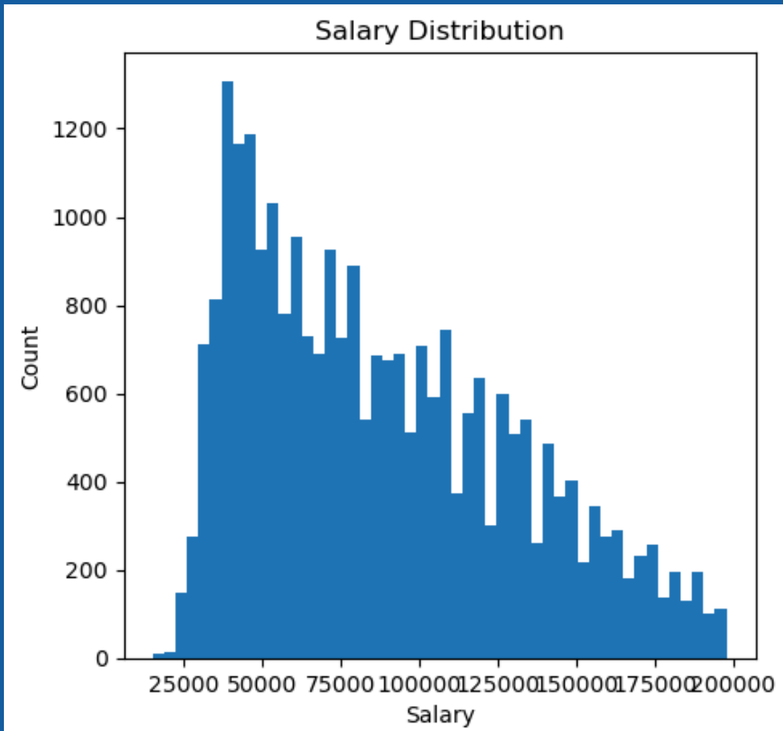
03

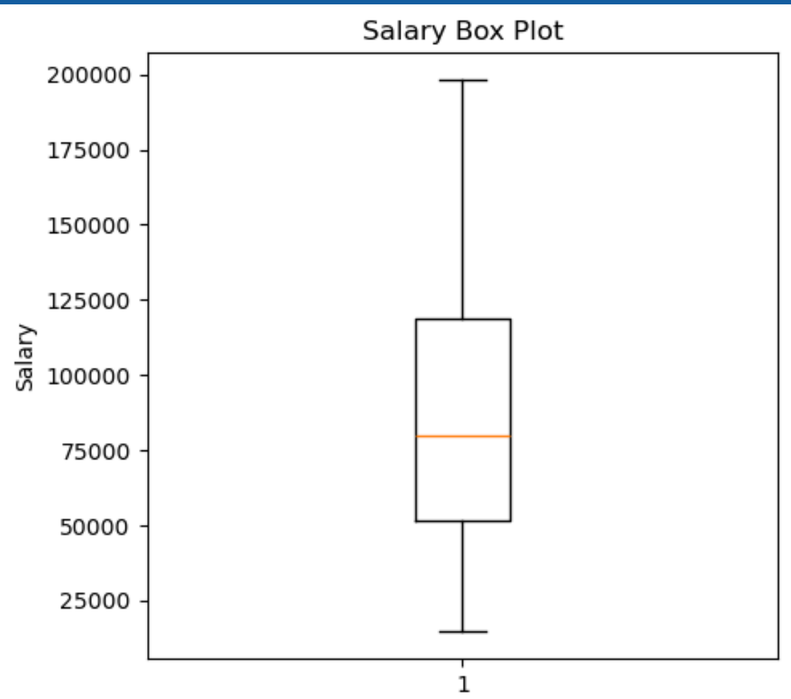
- Lowercase, remove special characters
- Tokenization, stopwords removal
- Context-Aware Lematization (POS tag)

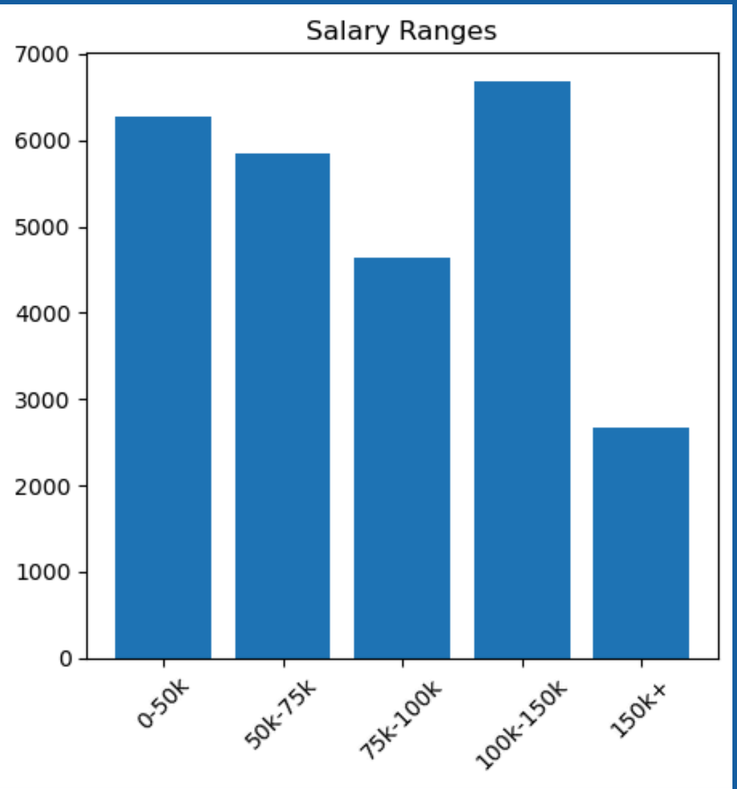
04

- Filtered the job description where the characters are above 1000
- Catch meaningful descriptions

EDA


Salary Distribution
Count
Salary
0
200
400
600
800
1000
1200
25000
50000
75000
100000
125000
150000
175000
200000


Salary Box Plot
Salary
25000
50000
75000
100000
125000
150000
175000
200000
1


Salary Ranges
0
1000
2000
3000
4000
5000
6000
7000
0-50k
50k-75k
75k-100k
100k-150k
150k+

MODELLING APPROACH

"How does feature extraction from job descriptions influence salary prediction accuracy?"
"Does a deep learning approach outperform traditional methods in NLP-based salary prediction?"

Basic Machine Learning Model

- Predicting salary using structured data features
- Variables: title, location, company, industry, experience level
- Random Forest
- R squared = 0.525

Bag-of-Words & TF-IDF

- Bag-of-Words (BoW): Identify the most demanded skill sets
- TF-IDF: A refinement of BoW that down-weights common words and up-weights rare but important words

BERT + Deep Learning

- BERT Embeddings: generates dense, contextualized embeddings
- Encodes full job descriptions into high-dimensional vectors that capture rich semantic meaning
- Deep Learning

TF-IDF + RF

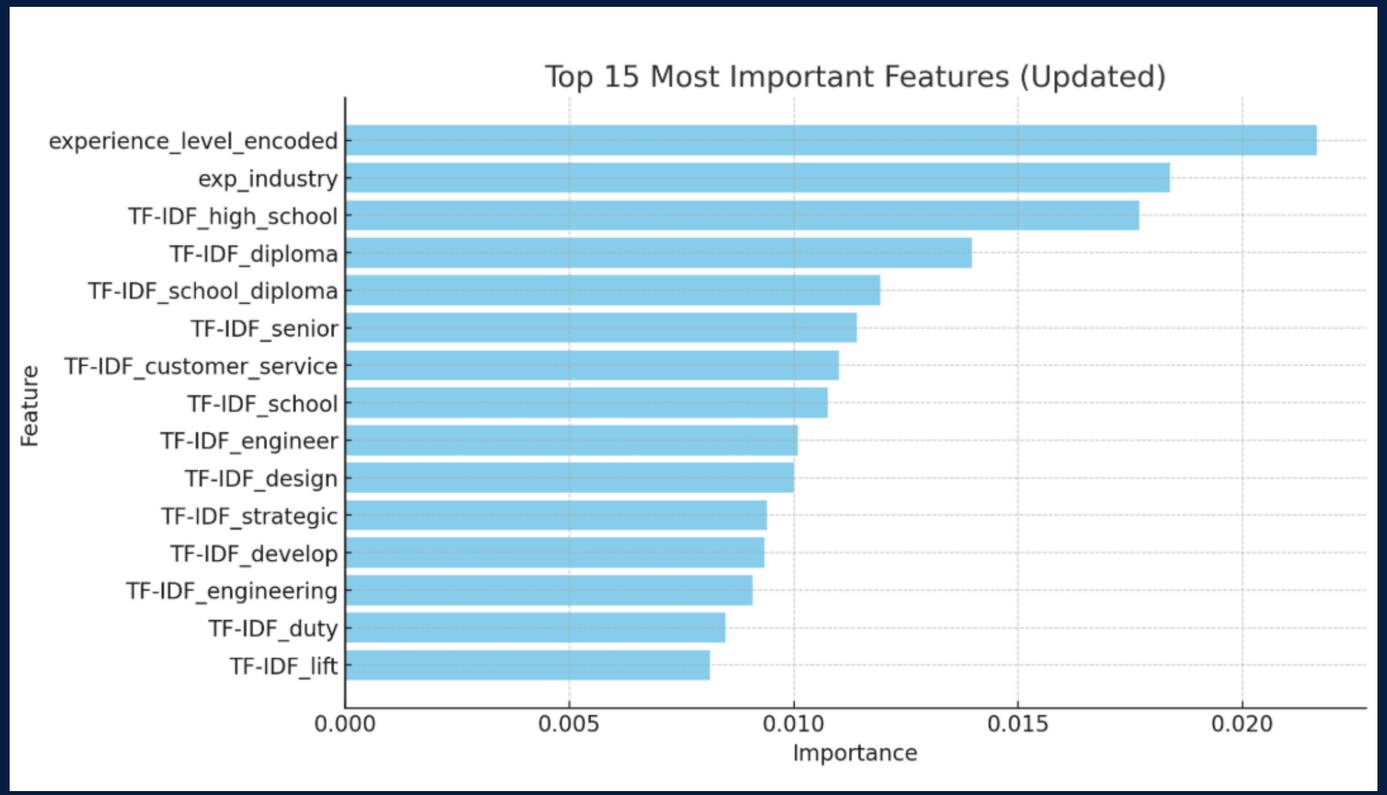
- R-squared = 0.592

- Hyperparameter tuning
  - Grid Search
    - Min_df = 0.01
    - Max_df = 0.9
    - Ngram_range = (1,3)

- Insights:
  - Experience level being the highest impact

  - Education-related keywords
    - high school
    - diploma

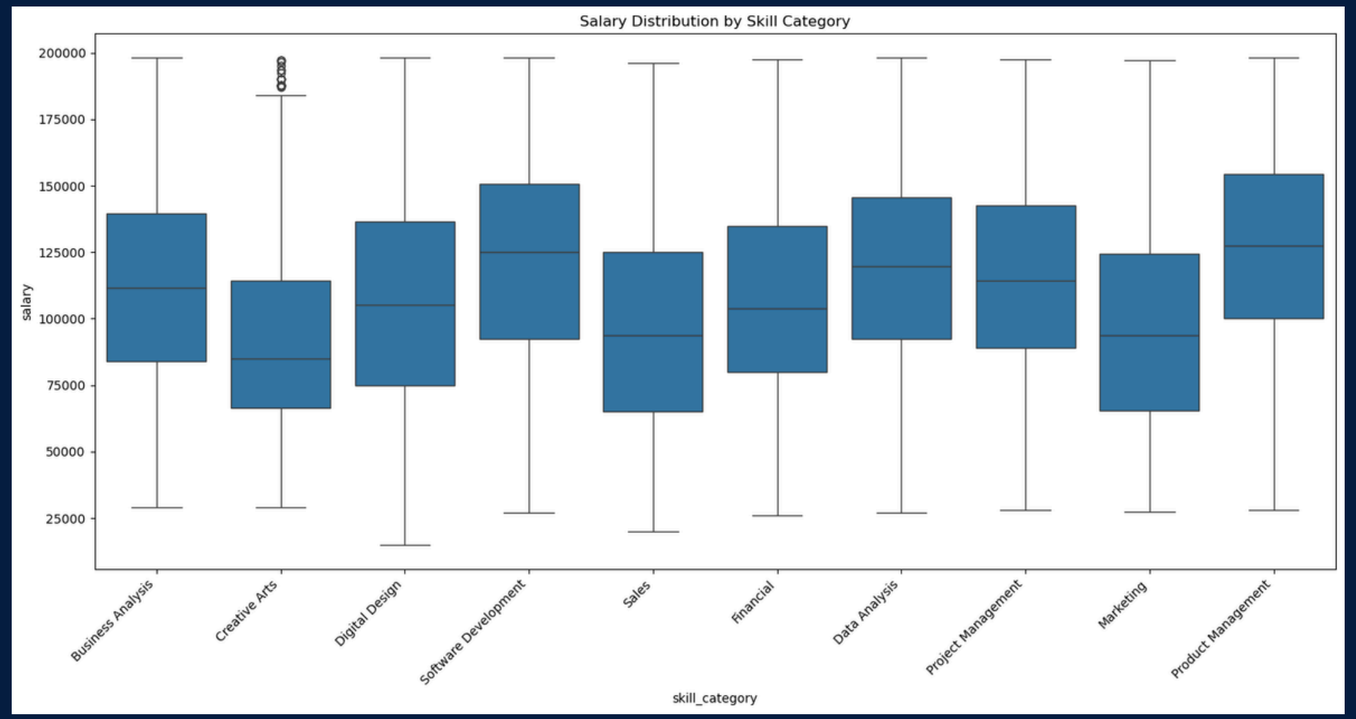
  - Job-specific terms
    - engineering
    - design
    - strategy



BAG OF WORDS

- **Skill Category Setup:**
  - Created 10 skill categories (e.g., Product Management, Data Analysis, Software Development) using comprehensive keyword lists (e.g., “product roadmap,” “sql,” “java”).

- **Skill Extraction:**
  - Extracted skill keywords found in each job description and mapped them to corresponding categories for encoding



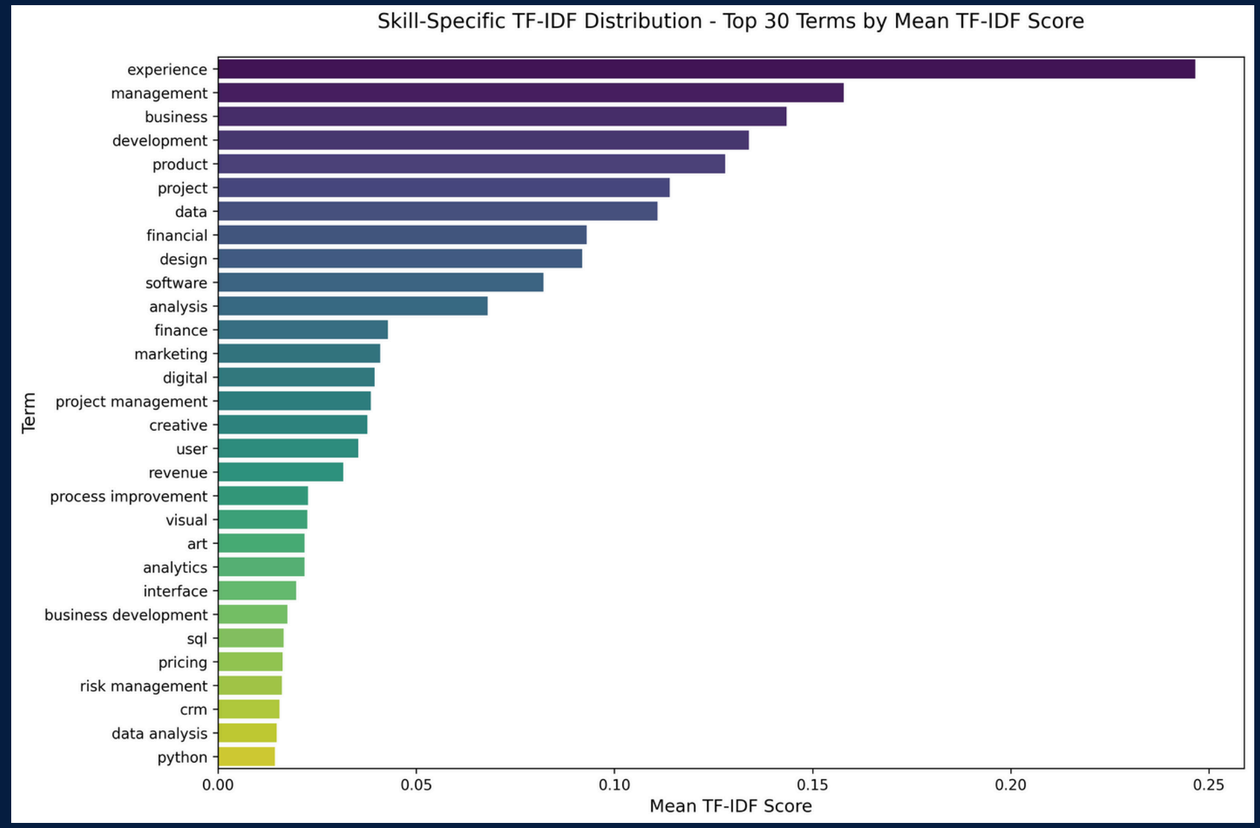
BOW + TF-IDF + RF

First Approach: Skill-specific TF-IDF

- R-squared = 0.623
- General TF-IDF parameter
  - Min_df = 0.01
  - Max_df = 0.9
  - Ngram_range = (1,3)
- Skill-specific TF-IDF parameter
  - Min_df = 0.03
  - Max_df = 0.8
  - Ngram_range = (1,3)

Second Approach: One hot encoding

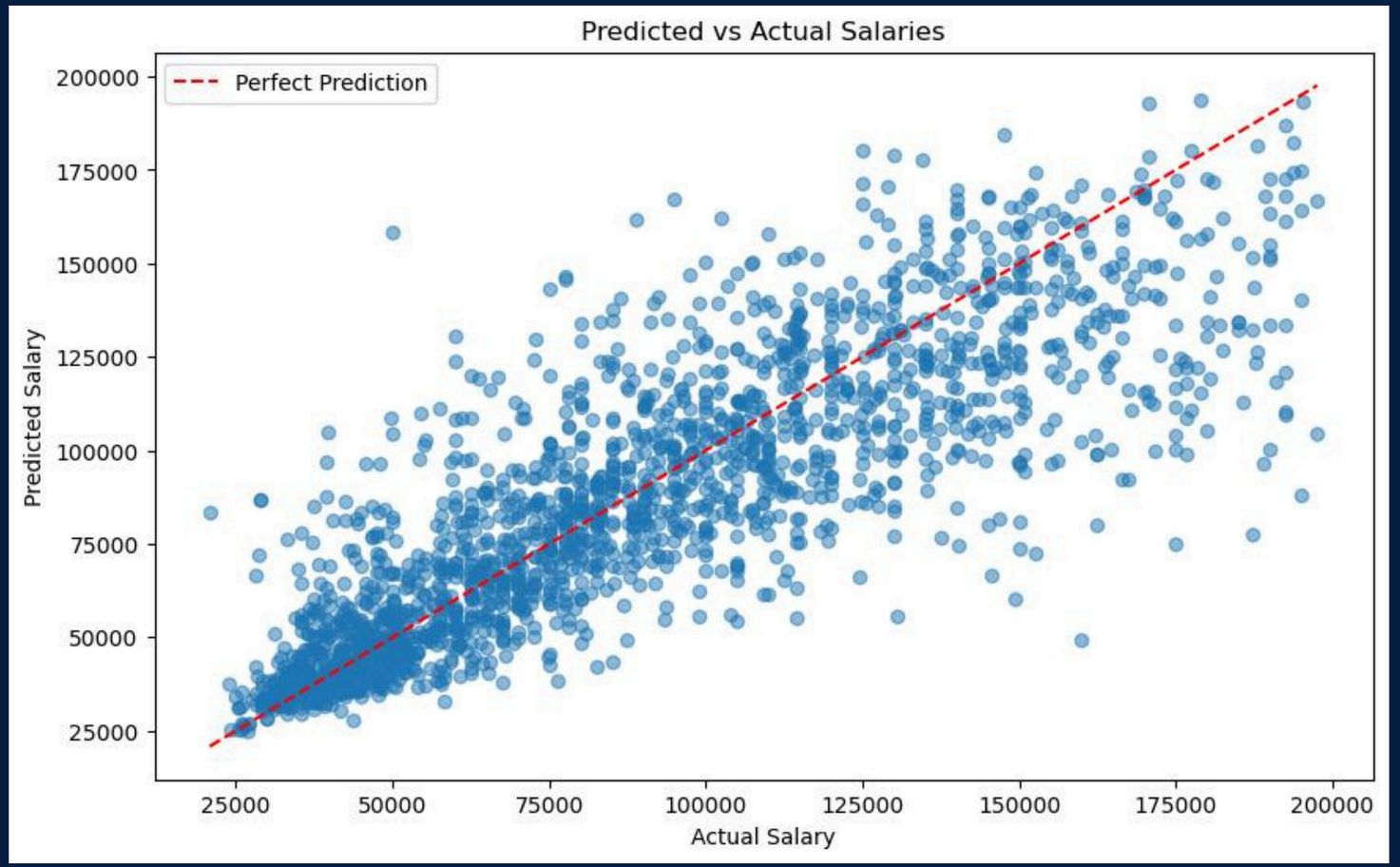
- R-squared = 0.647
- General TF-IDF parameter
  - Min_df = 0.01
  - Max_df = 0.9
  - Ngram_range = (1,3)



BOW + TF-IDF + NEURAL NETWORK

- R-squared = 0.732

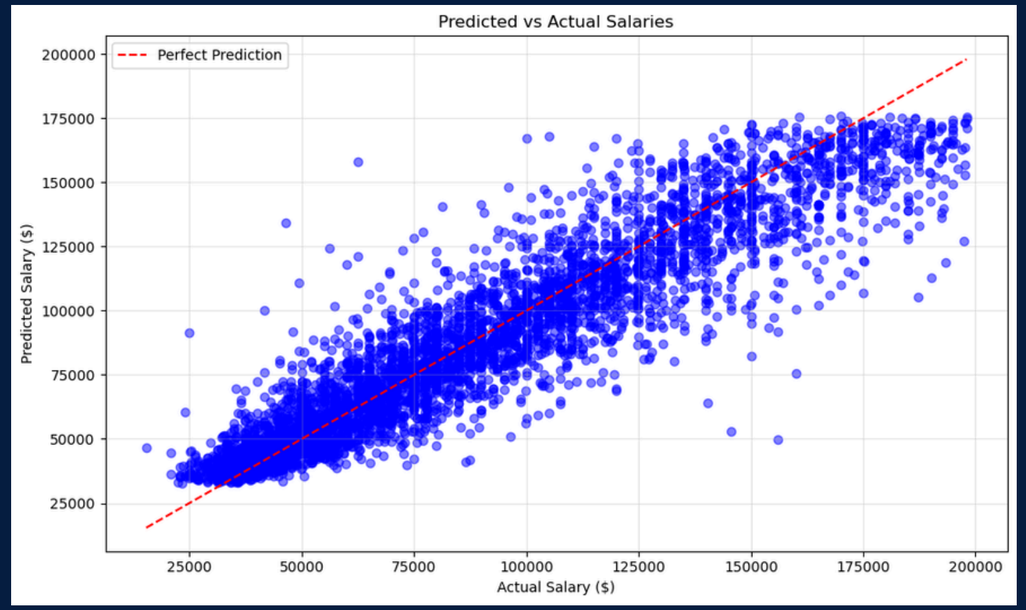
- General TF-IDF parameter
  - Min_df = 0.01
  - Max_df = 0.9
  - Ngram_range = (1,3)
- Skill-specific TF-IDF parameter
  - Min_df = 0.005
  - Max_df = 0.8
  - Ngram_range = (1,3)
- Neural Network parameter
  - hidden dims = 256, 128, 64
  - dropout rate = 0.1
  - learning rate = 0.0005
  - batch_size = 32

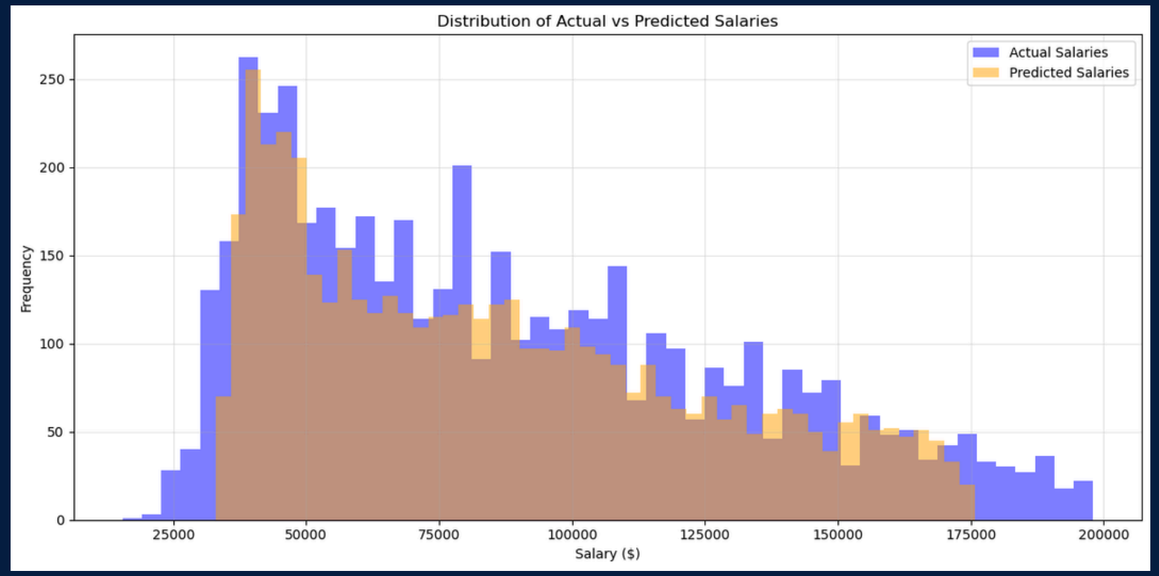


BERT + DEEP LEARNING

- Text Processing Tool:
  - BERT is an optimal choice due to its ability to model both short and long job descriptions while maintaining contextual awareness

- Model Performance Metrics
  - RMSE: $15,510.12
  - R2 Score: 0.8678





Model Performance

	R-Square	RMSE	Runtime
Basic Model	0.525	29,563.53	21 sec
TF-IDF + Random Forest	0.592	27,138.60	4 min 51 sec
BoW (encoding) + TF-IDF + RF	0.647	25247.65	44 min 42 sec
BoW + TF-IDF + Neural Network	0.732	21,797.11	9 min 50 sec
BERT + DL	0.868	15,510.12	540 min

BUSINESS VALUE

Helping Job Seekers Understand Industry Trends

- Job seekers can compare salaries across industries to identify where their skills are most valued
- Highlights which skill combinations lead to higher compensation, guiding career development and upskilling decisions

Enhancing Negotiation Power

- Candidates can negotiate confidently, backed by data on the expected salary for their role
- Tailored salary ranges based on location and employer profiles help in aligning negotiations to local markets

Building Long-Term Career Growth

- Guides job seekers in choosing industries with high salary growth potential.
- Informs candidates about additional skill sets to acquire for higher-paying roles.

Thank You