# Job Salary Prediction: A Text Analytics Approach

## Problem Statement

The job market suffers from significant salary information asymmetry, leaving job seekers at a disadvantage during compensation negotiations due to limited access to accurate benchmarks. Our project addresses this real-world problem by developing an advanced machine learning solution that predicts job salaries based on unstructured text data from job descriptions, leveraging NLP techniques.

## Data Processing & Feature Engineering

Our data processing began with a dataset of 27,900 LinkedIn job postings containing structured fields (company, industry, work type, company, experience level) and unstructured text (title, descriptions). We first applied basic text preprocessing including lowercase conversion, special character removal, tokenization, and stopword elimination. We also normalized salary ranges using US minimum wage ($15,000/year) as a lower bound and removed outliers through upper bound adjustment (Q3+IQR).

As our project progressed, we significantly enhanced our text preprocessing pipeline to improve feature quality. We implemented context-aware lemmatization with POS tagging to preserve semantic meaning while reducing vocabulary size. Additional to generic stopword lists, we also developed domain-specific stopword customization to remove non-informative words but have high TF-IDF score. We also implemented abbreviation handling to standardize industry terms (e.g., 'yrs' to 'years’), ensuring consistent representation of equivalent concepts.

For feature engineering, we developed a sophisticated approach to skill extraction by creating 10 distinct skill categories (e.g., Product Management, Data Analysis, Software Development) using comprehensive keyword lists. We extracted skill keywords from each job description and mapped them to corresponding categories, allowing us to capture domain-specific salary determinants that general text processing might miss.

## Modeling Strategy

Our modeling strategy involved systematically evaluating a progression of text analytics approaches, each building upon the insights from the previous. We began with a baseline Random Forest model using only structured data, then incorporated TF-IDF vectorization to capture term importance in job descriptions. We further enhanced our approach by implementing a sophisticated Bag-of-Words encoding combined with TF-IDF, which allowed us to represent domain-specific skills more effectively. Recognizing the potential of neural networks to capture complex relationships, we then combined our feature engineering with deep learning architectures, first using a standard configuration and later with enhanced text cleaning and optimized hyperparameters. For comparison, we also implemented a BERT-based approach with contextual embeddings. Each progressive refinement yielded substantial performance improvements, with our enhanced neural network striking an optimal balance between accuracy and computational efficiency.

## Model Performance

| Model | $R^2$ | RMSE | Runtime |
|---|---:|---:|---:|
| Baseline + RF (structured data only) | 0.515 | 29,620.21 | 40.7s |
| TF-IDF (General) + RF | 0.592 | 27,138.60 | 4m 51s |
| BoW + TF-IDF (General&Skill-Specific) + RF | 0.623 | 26,126.36 | 44m 42s |
| BoW + TF-IDF + Neural Network | 0.732 | 21,797.11 | 9m 50s |
| BoW + TF-IDF + NN (Enhanced Cleaning) | 0.829 | 17,502.23 | 5m 28s |
| BERT + Deep Learning | 0.868 | 15,510.12 | 540m |

Our model evolution demonstrated that while BERT embeddings achieve marginally better performance, our optimized neural network with enhanced preprocessing strikes an excellent balance between accuracy (0.829) and computational efficiency (5m 28s), making it more suitable for practical deployment.

## Discussion

Our approach incorporates several innovative technical contributions. The enhanced text cleaning pipeline preserved important salary-predictive terms while eliminating noise. This significantly improved the signal-to-noise ratio in our feature representation. We developed a novel skill-specific TF-IDF vectorization approach that separately processed general text and domain-specific terms with optimized parameters for each category. This allowed us to capture both broad contextual information and specialized technical requirements that impact compensation.

Through detailed feature importance analysis, we identified that experience level, industry expertise, and education consistently ranked as the strongest salary predictors across job categories. We also discovered that certain technical skills, particularly in software development and data analysis, commanded premium compensation regardless of industry sector. To ensure robust validation, we implemented a cross-validation strategy with stratification based on salary bands, ensuring reliable performance estimates across positions ranging from entry-level to executive.

## Conclusion

Our job salary prediction model demonstrates the power of advanced text analytics in addressing real-world challenges, achieving $R^2 = 0.8292$ through systematic refinement of preprocessing and modeling approaches. This dramatic improvement over baseline methods demonstrates how sophisticated NLP techniques with domain-specific feature engineering can extract valuable insights from unstructured job descriptions.

The model empowers job seekers with data-backed salary expectations aligned to their specific skills and locations, while offering career development guidance by identifying high-value skill combinations for strategic professional growth. By bringing greater transparency to job market compensation, our solution creates substantial economic value while promoting fairness in labor markets, making it suitable for practical deployment in recruitment platforms, career services, and compensation planning tools.