Salary Predictor For LinkedIn Job Postings

Zhichao Hao
Department of Computer Science
University of Toronto
Toronto, Ontario M5S2E4, Canada
jimmy.hao@mail.utoronto.ca

Jiatan Yuan
Department of Computer Science
University of Toronto
Toronto, Ontario M5S2E4, Canada
jiatan.yuan@mail.utoronto.ca

Xingjian Liu
Department of Computer Science
University of Toronto
Toronto, Ontario M5S2E4, Canada
xingjian.liu@mail.utoronto.ca

Zihan Wang
Department of Computer Science
University of Toronto
Toronto, Ontario M5S2E4, Canada
jameszh.wang@mail.utoronto.ca

Abstract

This study explores the application of machine learning techniques to predict salary levels from LinkedIn job postings, leveraging pre-trained models BERT and GPT-1 alongside a simpler MLP model for comparative analysis. By treating salary as discrete levels, this methedology aligns with common industry practices of listing salary ranges rather than exact figures, providing a more stable and interpretable framework for model evaluation. The results demonstrate the efficacy of BERT, benefiting from its bidirectional architecture, in capturing the nuanced context of job descriptions, which proves crucial for accurate salary prediction. (code: GitHub Link)

1 Introduction

Text classification is one of the most important field in natural language processing (NLP). This research utilizes LLMs to develop text classification models that estimate salary level based on online job postings. Accurate salary prediction helps create transparency in the labor market, providing crucial information for job seekers and employers. To date, few studies have addressed this application (Matbouli and Alghamdi [2022], Jackman and Reid [2013]), and while various machine learning strategies have been evaluated, none have utilized pre-trained large language models (LLMs) for this task. To advance research in this area of application, we develop two models that each incorporate BERT and GPT-1 to classify job posting data by the range of salary it falls into. We also develop a relatively straightforward MLP model for comparison. The overarching goal is to 1) develop models for predicting salary based on information from job postings and 2) identify whether BERT is indeed more suitable for classification tasks due to its bidirectional nature.

Rationale for Using Machine Learning The choice of machine learning is dictated by its unmatched capability in handling large and complex datasets. This feature is critical for the project due to several factors:

- **Handling Diverse and Extensive Job Data:** ML algorithms are particularly proficient in analyzing a wide variety of job descriptions across multiple sectors. This ability allows the model to identify subtle patterns and trends that span different industries, job levels, and geographic regions. It is this nuanced understanding that enables the accommodation of the complex variables affecting salary determinations.

Preprint. Under review.

- **Dynamic Adaptation and Learning:** Machine learning models continuously evolve by integrating new data, which allows them to refine and update their predictions. This dynamic learning is essential for maintaining the relevance and accuracy of the salary estimates, particularly in a job market characterized by rapid changes. Regular updates with new job postings and salary data ensure that the predictions remain reflective of the latest market conditions.
- **Providing Actionable Insights:** The application of ML not only supports the extraction of complex data patterns but also translates these into actionable insights for users. This dual capability facilitates informed decision-making for job seekers and helps employers design compensation strategies that are both competitive and equitable.

2 Background

2.1 Generative Pre-Training (GPT-1)

Generative Pre-Training (GPT-1)Radford et al. [2018] is a novel approach that introduces pre-training a language model on a diverse corpus of text followed by fine-tuning on specific downstream tasks. The GPT-1 model utilizes the Transformer architecture, emphasizing the learning of deep, bidirectional representations from unlabeled text. By pre-training on an extensive range of text, the model captures the intricacies of language, enabling it to excel at various language understanding tasks. This pre-training allows the model to generalize across domains and tasks, facilitating rapid adaptation to new language tasks with minimal task-specific adjustments.

2.2 Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT)Devlin et al. [2018] is a groundbreaking technique that fundamentally changes the way pre-training and fine-tuning are conducted in language models. Unlike previous models, BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right contexts in all layers. The architecture enables the model to learn a more nuanced understanding of language context and structure. The BERT framework can be fine-tuned with just an additional output layer, allowing the model to set new state-of-the-art benchmarks for a wide range of natural language processing tasks, including but not limited to question answering and natural language inference.

3 Related Work

3.1 Statistical Machine Learning Regression Models for Salary Prediction

This study develops a holistic framework for predicting salaries across all job titles in the Saudi Arabian economy using machine learning (ML). It compares the performance of five ML algorithms on survey data to estimate annual salaries based on occupational features and organizational characteristics. The Bayesian Gaussian process regression significantly outperformed multiple linear regression in predicting salaries across economic activities, while artificial neural networks were most effective for predicting salaries across major occupational groupsMatbouli and Alghamdi [2022].

3.2 PayPredict

PayPredict is a Chrome extension that offers real-time salary estimates on LinkedIn profiles for recruiters, HR professionals, and hiring managers. Utilizing data from Figures, it provides accurate salary insights based on job title, experience, and locationFigures [2024].

4 Data Description and Processing

4.1 Dataset Overview

We utilize the LinkedIn Job Postings - 2023 dataset available on Kaggle. This dataset comprises over 33,000 job postings from LinkedIn, each described by 27 attributesArshkon [2023]. Our dataset is

also supplemented by an additional 30,000 job postings data retrieved from LinkedIn, Glassdoor, Indeed and Zip Recruiter, although it should be noted that this dataset may not capture all sporadic postings and there exists missing values for some fields. Nonetheless, it makes the dataset more representative of the job postings on the web.

4.2 Selected Attributes and Justification

We selected 4 categorical attributes and 3 semantic attributes for the purpose of our study. The categorical attributes include: *industry*, *work_type*, *location*, *experience_level*. The textual attributes include *company_name*, *position_title*, *description*. These attributes were chosen because we believe they have more direct relevance to salary predictions and job categorization compared to other attributes. Attributes such as *views* and *sponsored* are excluded as they are less indicative of intrinsic job characteristics.

4.3 Data Cleaning Steps

We process the original data with the following steps:

- **Filtering Incomplete Data:** We filter out entries that lack salary data, reducing our dataset to approximately 30,000 samples. Then, we also filtered out entries with NA values in [*formatted_experience_level*, *work_type*, *industry*] to focus on jobs only with sufficient data, resulting in a dataset with around 14,000 samples.
- **Salary Data Standardization:** Entries lacking *med_salary* are processed by calculating the average of *max_salary* and *min_salary*, while entries with *med_salary* are used as is. All salaries are standardized to annual figures based on *pay_period*. Although the dataset does not specify the currency of the salary figures, due to the maintainer's location in the United States, we assume all salary figures are in U.S. dollars (USD). This assumption should be considered when interpreting the salary data, especially for applications outside the U.S.
- **Categorize into Salary Levels:** The salary values in our data predominantly ranges from 10K to 160K, thus we decided to categorize the salaries into the following levels: "10K and below", "10K to 11K", "11K to 12K", "13K to 14K", ..., up to "160K and above".
- **Location Cleaning:** Only the state abbreviation is retained from the location data to examine the impact of geographic location on salary, which will be evaluated for its predictive relevance before final inclusion.
- **Job Title Cleaning:** We use Spacy for tokenization to simplify job titles. A frequency-based vocabulary will be created to facilitate embedding, representing the presence of words in titles.
- **Description Cleaning:** Descriptions are processed using Spacy to remove all special and space characters, streamlining the text for more effective natural language processing. This step improves the quality of text data, ensuring that our models focus on meaningful content without noise.

These steps aim to refine the dataset for efficient and effective analysis, ensuring that the data feeding into our models is accurate and representative of the variables most critical to the study's outcomes.

The processed dataset is split into training, validation, and test dataset each contains 10000, 3000, and 1000 samples accordingly.

5 Baseline Model

We use a simple Multi-Layer Perceptron (MLP) for our baseline model. This model is trained to predict the salary category based on the categorical features of the job postings.

5.1 Model Input

The baseline MLP model takes a concatenation of one-hot encoded vectors, representing categorical features of job postings as input. These features include *work_type*, *industry*, *formatted_experience_level*, and *location*. Each category is transformed into a binary vector where only

the index corresponding to the specific category is marked as one, while the rest are zeroes. All these vectors are then flattened and concatenated to form a single input tensor that encodes the entire categorical data profile for a job posting.

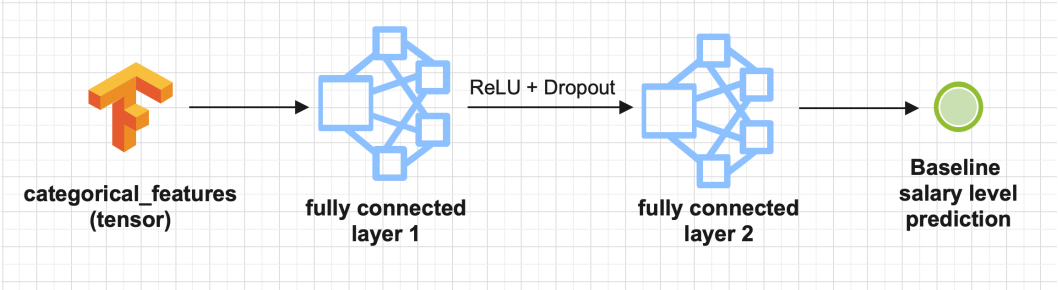

Figure 1: Architecture of the Baseline MLP Model

5.2 Model Architecture

- **Hidden Layer:** The input tensor is first processed by a hidden layer comprised of a specified number of neurons. This layer uses ReLU (Rectified Linear Unit) activation to introduce non-linearity into the model, allowing it to learn complex patterns in the data.
- **Dropout Regularization:** To mitigate the risk of overfitting, a dropout layer is applied after the ReLU activation with a rate of 0.1, randomly zeroing out a portion of the layer's outputs during training.
- **Output Layer:** The final output is computed using a fully connected layer that maps the hidden layer's representations to a single continuous salary value.

5.3 Model Output

The output of the baseline MLP model are logits that represent the likelihood for the input job posting to belong to each category. This can then be evaluated using the argmax function to retrieve the mostly likely salary range of the input job posting.

5.4 Training and Evaluation

The model is trained using Cross-Entropy Loss to optimize the weight parameters. The SGD optimizer is chosen for its computational and memory efficiency. Model performance is evaluated based on the Cross-Entropy Loss and its classification accuracy on both the training and validation datasets.

This baseline model, while lacking the advanced natural language processing abilities of GPT-1 or BERT, provides a foundational benchmark for comparing the efficacy of various feature representations and neural network complexities in the task of salary prediction.

6 Architecture

Initially, our approach centered on a regression model (like in Matbouli and Alghamdi [2022]), targeting salary as a continuous numerical variable. However, the preliminary results proved to be highly volatile and challenging to assess accurately (See Figure 7 in Appendix). This instability led us to pivot towards a classification framework. By categorizing salary into discrete levels—as detailed in the data processing section—we could more effectively stabilize model outputs and improve evaluative metrics. This classification approach aligns with the typical format of job postings on job posting sites like LinkedIn, which generally list salary as a range rather than a precise figure.

We propose three new models that classifies job posting data according to salary ranges. The first is a slight improvement over the baseline model that incorporates title embedding to capture useful information from the job title. The latter two each incorporates a different large language model - BERT and GPT-1 - to process textual data. BERT, with its bidirectional training architecture, is hypothesized to be particularly well-suited for tasks that benefit from a deep understanding of

language context. GPT-1, though unidirectional, is considered for its efficacy in handling large-scale language models. The two models make up a fair comparison due to their similarity in size, both having around 110 million parameters.

All three models are aimed at deriving salary predictions from textual data contained in job postings. The architectures share a common approach to handling categorical data but employ different strategies to process the textual content.

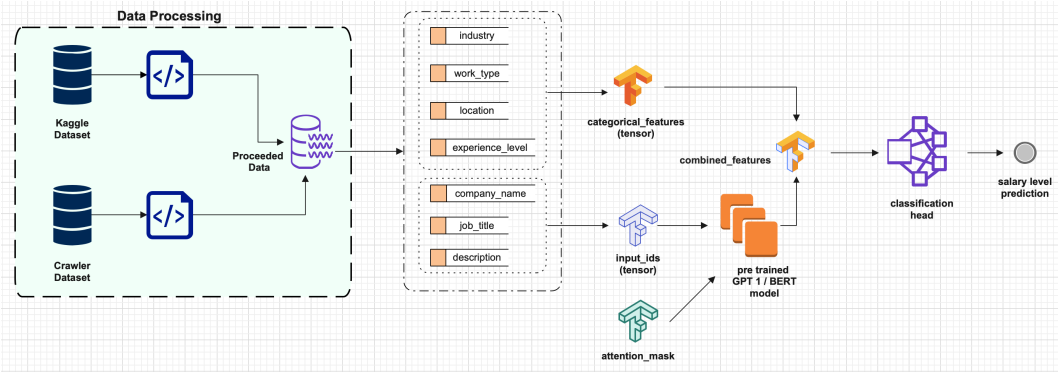

Figure 2: Architecture of Primary Models

6.1 MLP + Title Embedding Model Architecture

For this architecture, we generate a vocabulary for the *job_title* field according to the appearance frequency of words. Then we create an embedding for each job entry using this vocabulary table. It is composed of three main components:

- **Job Title Embedding:** We generate a very basic embedding of job titles using our vocabulary table crafted according to word frequency.
- **Categorical Feature Encoding:** Categorical features including *location*, *formatted_experience_level*, *industry*, and *work_type* are encoded using one-hot encoding.
- **Feature Concatenation and Classification:** The job title embedding are concatenated with the one-hot encoded categorical features. This combined feature vector then passes through an classification head with the same structure as our baseline model for salary prediction.

6.2 GPT-1/BERT Based Model Architecture

This architecture utilizes OpenAI GPT-1 and Google BERT to process textual features such as *company_name*, *job_title*, and *job_description*. During training, the LLM module is **fine-tuned** along with the rest of the model (the classification head). It is composed of three main components:

- **Textual Data Processing:** Values belonging to the fields *company_name*, *job_title* and *job_description* is concatenated into a single string, separated by space and truncated to 512 tokens in length. The concatenated string is fed into the GPT-1 model. We then perform average pooling on the tokens in the final hidden state of the output to obtain the embedding for the textual data.
- **Categorical Feature Encoding:** Categorical features including *location*, *formatted_experience_level*, *industry*, and *work_type* are encoded using one-hot encoding.
- **Feature Concatenation and Classification:** The textual data's embedding is concatenated with the one-hot encoded categorical features. This combined feature vector then passes through an classification head with the same structure as our baseline model for salary prediction.

7 Results

7.1 Regression Model

As we previously mentioned, we started by training regression models to directly predict salary as a number. However, the accuracy of this approach is nowhere near satisfying. Although we experimented with multiple hyperparameters, the models still underfit the data. We got extremely large Mean Absolute Error (MAE) on both training, validation, and testing datasets. Models that integrated BERT or GPT behaved similarly. (Appendix.Figure 7)

7.2 Classification Model

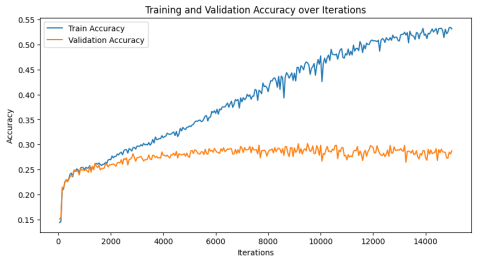

Figure 3: Training and Validation Accuracy for Baseline MLP Model

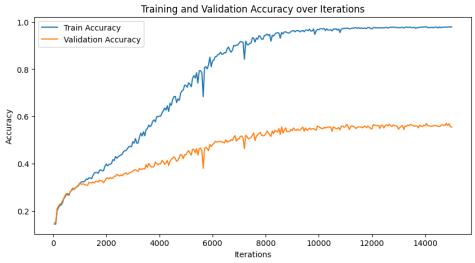

Figure 4: Training and Validation Accuracy for MLP with Title Embedding Model

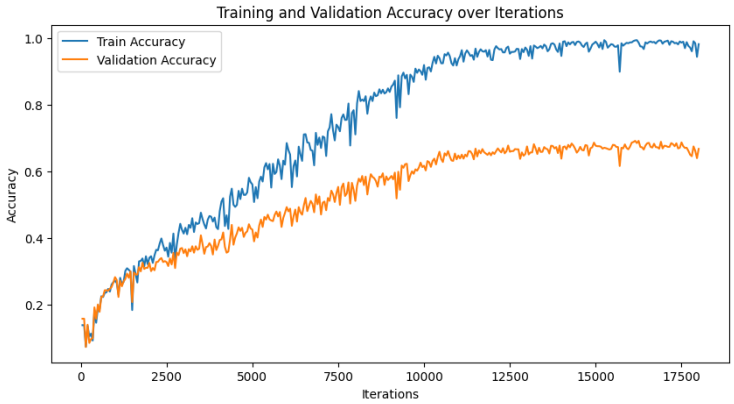

Figure 5: Training and Validation Accuracy for Fine-tuned BERT Model

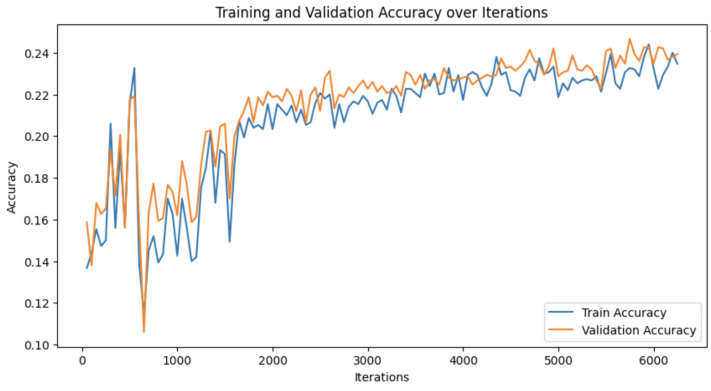

Figure 6: Training and Validation Accuracy for Fine-tuned GPT-1 Model

Model	Test Accuracy (%)
Baseline MLP	28.76
MLP with title embedding	55.46
Fine-tuned BERT	**67.42**
Fine-tuned GPT1	23.34

Table 1: Accuracy of Models in Predicting Salary

The results improved significantly after switching to classification. The baseline MLP without title_emb converged with am accuracy of 28.76% on the testing set, and the MLP using title_emb converged with test accuracy of 55.46%. The BERT-based model achieves the highest result out of the three, at 67.42% accuracy. The GPT model isn't performing as we expected, we encountered sudden drops in accuracy during training and were not able to resolve this issue. Due to page limitations, we only include the training curve of the Fine-tuned BERT model here, images for the other models can be found in the Appendix.

Overall, we found that the information captured from the title using the title embedding is very helpful for classifying salary efficiently. While the Bert model has gained much better accuracy, both the scale and the training time of the Bert model are much higher at the same time.

8 Discussion

8.1 Challenges Encountered

When the model predicts a salary range based on the data, there is a discrepancy between the predicted range and the actual salary range provided in a job listing. For example, if a job posting lists a salary range from $70K to $100K, the median (and thus the target for your model) would be $85K. However, if the model predicts a salary range of $70K to $80K, it is technically correct in that it captures part of the actual range. Yet, according to your standardized target of $80K to $90K (centered around the median), this prediction would be off-target.

8.2 Practical Implications

The study's findings are particularly relevant for platforms like LinkedIn, where job seekers and employers benefit from greater transparency in salary expectations. By automating salary range predictions, our models can help reduce information asymmetry in the job market, aiding in more equitable job negotiations.

8.3 Future Research Directions

As mentioned, there is a potential misalignment between how salary ranges are processed and interpreted in the model versus how they are presented in real-world job listings. A more precise standardization method to divide salary level may significantly improve the performance.

Further research could also explore the integration of additional contextual features, such as company size and posts, which could enhance the models' predictive accuracy. Additionally, exploring newer models that incorporate both bidirectional and generative capabilities could provide further advancements in this field.

8.4 Conclusion

Our research demonstrates the potential of using pre-trained models to classify salary ranges from textual job data effectively. BERT, in particular, proved to be well-suited for this task, leveraging its bidirectional training to achieve high levels of accuracy. The transition from a regression model to a classification approach was pivotal, addressing the inherent challenges of salary prediction and aligning more closely with industry standards.

The outcomes of this study not only enhance our understanding of the capabilities of different neural network architectures in handling real-world data but also pave the way for more sophisticated

approaches to salary prediction. Ultimately, this research present a reasonable approach for future systems aimed at providing real-time, accurate salary information to job seekers and employers alike.

9 Ethical Considerations

Data Privacy and Consent Issues: Utilizing the LinkedIn Job Postings - 2023 dataset introduces significant ethical challenges, particularly concerning privacy. Although the dataset consists of publicly available job postings, the public nature of this data does not automatically equate to consent for its use in this analysis. Individuals and companies featured in the dataset might not be aware that their information is being used for salary prediction, raising issues about informed consent and the ethical use of publicly scraped data.

Risk of Bias and Representational Fairness: The dataset may also contain inherent biases that do not accurately reflect the broader job market. Such biases could be due to the overrepresentation or underrepresentation of certain job types, industries, or geographic locations. Relying on this dataset for salary predictions might inadvertently amplify existing workplace inequalities. For instance, if certain high-paying industries are overrepresented, the model could skew salary estimations upwards, misleading job seekers and policy makers.

Model Limitations and Potential for Bias Amplification: The predictive model, employing KNN regression and text embeddings derived from job titles and descriptions, is particularly susceptible to inheriting and potentially amplifying any biases present in the data. Since the model primarily relies on job titles and descriptions to estimate salaries, it may perpetuate existing biases linked to how different jobs are valued or perceived in various cultural or socio-economic contexts. This approach risks producing unfair predictions and maintaining salary disparities across different demographic groups or job types. The model's accuracy and fairness are thus contingent on the representativeness and impartiality of the data it is trained on.

10 Description of Individual Contribution

- Zhichao Hao: Main developer of Data Processing; Debug GPT-1 Model; Draw Architecture Diagram; Write Abstract, Introduction, and Disscussion, draft Data Cleaning Step and Architecture.
- Jiatan Yuan: Main writer of the paper, write the Abstract, Rationale for Using Machine Learning, Background, Related Work, Data Description and Processing, Architecture, Ethical Consideration.
- Xingjian Liu: Main developer of GPT-1 and MLP Model; wrote the Crawler; Draft GPT-1 and MLP Architecture. Improved the paper's introduction and architecture.
- Zihan Wang: Main developer of BERT Model; Converted Models into classification models; Fine-tune every model; Draft Data Cleaning Step and BERT Architecture.

References

Arshkon. Linkedin job postings - 2023. `https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/data`, 2023. Accessed: April 18, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Available: `https://arxiv.org/pdf/1810.04805.pdf`.

Figures. PayPredict by Figures - Chrome Web Store. `https://chrome.google.com/webstore/detail/paypredict-by-figures/ohogbolcbnnmagamkjffiadkagfoghph`, 2024. Accessed: Feb. 27, 2024.

Shaun Jackman and Graham Reid. Predicting job salaries from text descriptions, Apr 2013. URL `https://open.library.ubc.ca/collections/graduateresearch/42591/items/1.0075767`.

Y.T. Matbouli and S.M. Alghamdi. Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations. *Information*, 13(495), 2022. doi: 10.3390/info13100495. URL https://doi.org/10.3390/info13100495.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Appendix

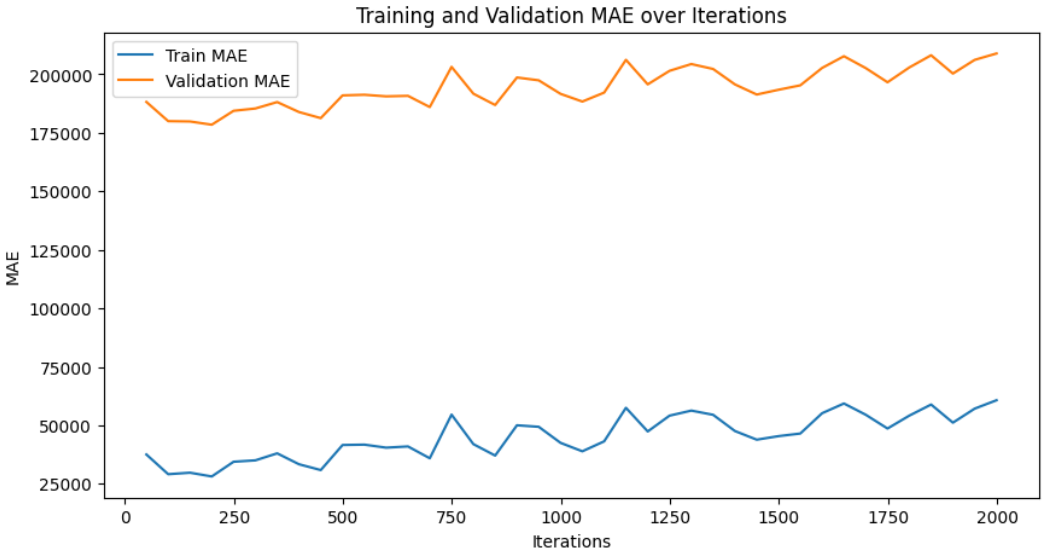

Figure 7: Training and Validation MAE for MLP Regression Model