SCHOLARLY REVIEW

SCHOLARLY REVIEW **JOURNAL**

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

Salary prediction using machine learning

By Michael Xu

AUTHOR BIO

Michael Xu is a Del Norte High School student with a passion for computer science and math. They have conducted research on machine learning and are particularly interested in this field. Outside of academics, Michael Xu is involved in game development and coding. They plan to pursue computer science and continue exploring applications of machine learning. In their free time, they enjoy chess and music.

ABSTRACT

This research explores how individual and demographic factors, such as professional experience, location, and education, influence salary trends and aims to uncover insights into their relative importance. Accurately predicting salaries helps understand trends in the workforce and assists career planning. Machine learning offers powerful tools for analyzing complex relationships between factors such as basic demographics, and wages. A synthetic dataset of 1,000 samples with 6 features is used. The performance of three neural network models is evaluated through test loss values, with the model incorporating a single hidden layer achieving the lowest loss of 0.1415. The model with no hidden layer and 2 hidden layers recorded test losses of 0.1435 and 0.145 respectively. Among the features, high school education, PhD, and years of experience demonstrate high contributions to salary predictions, as measured by permutation importance. This study demonstrates the efficiency of neural networks in predicting salaries, even with a synthetic dataset, showcasing their ability to generalize well and outperform manual prediction methods. Beyond prediction accuracy, interpretability techniques like permutation importance provide valuable insights into determinants of salaries for job seekers and employers. This study reinforces the effectiveness of machine learning in salary prediction.

Keywords*:* *Machine learning*, *neural network*, *salary*, *artificial intelligence*, *career planning*, *predictive model*, *training*, *synthetic data*

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

INTRODUCTION

In today's rapidly evolving job market, identifying major salary influencers remains a challenge, often affected by a variety of factors such as professional experience, location, and education. These factors often interact in unpredictable ways, making it difficult to identify their true impact. This complexity is further highlighted in McKinney et al. (2022), which shows that earnings differences are driven not only by measurable factors like hours worked, geographic division, and education but also by unexplained variations, particularly in higher income brackets. The unpredictability and nuanced interplay of these elements shows the difficulty in accurately predicting salaries.

Machine learning has emerged as a powerful tool to analyze complex datasets and discover hidden relationships. It uses a large amount of data and makes use of intricate algorithms to model the non-linearity between variables. Among the different types of machine learning models, the neural network model stands out as being able to capture complex relationships and being able to adapt to a variety of data structures, making them a versatile choice for data analysis.

This study uses machine learning, specifically neural networks to better capture the complex relationships between factors. While many other researchers have applied machine learning to this problem, this study serves as a practical exercise to evaluate a neural network on salary prediction. For the purpose of this study, a synthetic dataset available on Kaggle was used to simulate real-world salary data, providing a controlled environment for model training and evaluation ("Salary prediction data," n.d.). Consequently, this research focuses on the practice of building and tuning a model.

METHODOLOGIES

Dataset overview

The dataset utilized in this study contains 1,000 samples with six features and one target label. Each feature provides descriptive information about an individual, including:

- *Level of education*: Refers to the highest degree achieved by the individual, categorized as High School, Bachelor's, Masters, or PhD.
- *Geographic location*: Refers to whether the individual resides in an urban, suburban, or rural area.
- *Job title*: Refers to the career designation, which includes Analyst, Director, Engineer, or Manager.
- *Years of experience*: Refers to the number of years the individual has worked in their field.
- *Age*: Refers to the age of the individual.
- *Gender*: Refers to the individual's gender as either male or female.

The target label is Salary, which represents the annual income of the individual. In supervised learning, we use the information provided by the aforementioned features to predict the label, or the salary in this case. A sample of the unprocessed dataset is shown in Table 1.

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

Table 1
Two dataset samples

Education	Experience	Location	Job_Title	Age	Gender	Salary
High_School	8	Urban	Manager	63	Male	84620.053665
PhD	11	Suburban	Director	59	Male	142591.255894

Dataset encoding

The dataset used in this study is clean, containing no missing data, duplicate entries, or inaccurate values. Therefore, no data cleansing is required at this stage.

Several features in the dataset: education, location, job title, and gender are categorical features. To convert these categorical values into a format suitable for machine learning models, one-hot encoding is applied. This technique transforms each category into a binary number, where each category is represented as a separate feature with values of either 0 or 1. The resulting one-hot encoded columns include the following:

- *Education*: Education_High School, Education_Bachelor, Education_Master, Education_PhD
- *Location*: Location_Rural, Location_Suburban, Location_Urban
- *Job Title*: Job_Title_Analyst, Job_Title_Director, Job_Title_Engineer, Job_Title_Manager
- *Gender*: Gender_Male

For example, instead of having a single column with values such as "High School" or "PhD", each education level is represented as a separate column. A value of 1.0 in a column indicates that the individual belongs to that category, while the remaining columns for that feature will have 0.0. See Table 2 for a demonstration of the encoding of the Education feature.

Table 2
Two samples for encoded education feature

Education_Highschool	Education_Bachelor	Education_Master	Education_PhD
1.0	0.0	0.0	0.0

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

Education_Highschool	Education_Bachelor	Education_Master	Education_PhD
0.0	0.0	0.0	1.0

Since the feature gender only consists of two possible values in this dataset, instead of creating two features (Gender_Male and Gender_Female), a singular feature (Gender) can fully encapsulate the data, where 1 represents male and 0 represents female.

Dataset splitting

The dataset is partitioned into two subsets to facilitate model development and evaluation. Specifically, 80% of the dataset is allocated for training, where the model learns and optimizes its parameters, while the remaining 20% is reserved for testing, providing an unbiased assessment of the model's predictive capabilities. A constant random state is used during this process to ensure the results are reproducible. This approach balances the need for sufficient training data with the importance of having a reliable evaluation set.

Feature Scaling

To ensure that features contribute proportionally and equally to the model's performance, numerical data in the dataset is standardized. Standardization transforms the features to have a mean of 0 and a standard deviation of 1, which improves the convergence and stability of the training process by ensuring that features with larger values do not dominate those with smaller values.

Particularly, this is calculated using the following formula to find each individual value:
x'=x - ,
where:

- x' is the scaled value
- x is the original value
- is the mean of the feature in the training set
- is the standard deviation of the feature in the training set

Continuous variables such as years of experience and age were standardized while categorical features resulting from one-hot encoding were left unscaled, as scaling them is unmeaningful and distorts their binary representation.

Separate scalers were fitted on the training set to prevent data leakage and applied to both the training and testing sets. This ensures that the model's evaluation of the test set remains unbiased. The finalized preprocessed data is shown in Table 3.

Table 3

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

Two preprocessed samples from the dataset

YE	Age	GM	JA	JD	JE	JM	LR	LS	LU	EB	EH	EM	EP	Salary
0.652 25572	1.219 3170 6	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.20556 321
1.498 94836	1.293 2038 6	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	-1.1555 6687

Note. Legend: YE: Years of Experience, GM: Gender_Male, JA: Job_Title_Analyst, JD: Job_Title_Director, JE: Job_Title_Engineer, JM: Job_Title_Manager, LR: Location_Rural, LS: Location_Suburban, LU: Location_Urban, EB: Education_Bachelor, EH: Education_High School, EM: Education_Master, EP: Education_PhD

MODEL DEVELOPMENT

Neural networks were chosen for this study due to their ability to model complex, non-linear relationships between features, something in which traditional regression models often fail to capture successfully. It consists of three key components: an input layer matching the number of features (13 in this case), hidden layers to introduce non-linearity and enhance learning, and a single-node output layer to predict the salary. The model is optimized using the Adam optimizer and is evaluated with the Mean Squared Error (MSE) loss function. Mean Squared Error (MSE) was chosen as the loss function because it penalizes larger errors more heavily than Mean Absolute Error (MAE), making it suitable for salary prediction where large deviations from actual values are more impactful. Model training utilizes 80% of the dataset, as previously described, and is conducted over 200 epochs to balance between efficiency and prediction accuracy. The specific choice of zero, one, or two hidden layers was made in this study to test how model complexity affects predictive performance.

Model 1: No Hidden Layer – The model (see Figure 1) includes only an input and an output layer, making it equivalent to a linear regression model. Its simplicity confines it to capturing direct, proportional relationships between features and the target variable, making it ideal for datasets with predominantly linear trends.

Figure 1
Code snippet for Model

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

```
model1 = Sequential()
model1.add(Dense(1))
model1.compile(loss='mean_squared_error',
optimizer='adam')
```

Model 2: One Hidden Layer – This model (see Figure 2) can support non–linear relationships by introducing a single hidden layer with four nodes and a ReLU activation function. The hidden layer acts as an intermediate step that enables the model to transform the inputs, improving predictions for moderately complex patterns.

Figure 2
Code snippet for Model 2

```
model2 = Sequential()
model2.add(Dense(4, activation='relu'))
model2.add(Dense(1))
model2.compile(optimizer='adam', loss='MSE')
```

Model 3: Two Hidden Layers – This model (see Figure 3) with one hidden layer containing four nodes and another containing two nodes, both using ReLU activation, is tailored for datasets with intricate interactions among features. The layered structure facilitates a hierarchical learning process, where the first layer identifies simpler patterns, and the second layer combines these to detect more nuanced interactions. This design comes with a trade-off, as increased depth can also lead to overfitting on smaller datasets, like the one used in this study.

Figure 3
Code snippet for Model 3

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

```
model3 = Sequential()
model3.add(Dense(4, activation = 'relu'))
model3.add(Dense(2, activation = 'relu'))
model3.add(Dense(1))
model3.compile(optimizer='adam', loss='MSE')
```

Each model is trained for 200 epochs using the Adam optimizer. Training loss was recorded across epochs and plotted to observe convergence behavior. Test loss was computed to evaluate performance.

Permutation importance was used to quantify each feature's contribution to the model's predictions. This method provides insights into which features most strongly influence the target variable. The feature importances are visualized as a horizontal bar chart, ranking features by their contribution to prediction accuracy.

RESULTS

The performance of the three neural network models is evaluated using test loss values and visual analyses. Each model shows unique characteristics, reflecting its complexity and capacity to capture patterns within the dataset.

MODEL PERFORMANCE

Model 1: The simplest model, with no hidden layers, recorded a test loss of **0.1435**. The test loss reflects its linear pattern and limited capacity.

Model 2: Featuring a single hidden layer with ReLU activation, this model achieved the best test loss of **0.1415** and more effectively captures the relationship.

Model 3: With two hidden layers, Model 3 obtained a test loss of **0.145**. Its increased complexity resembled slight signs of overfitting.

TEST LOSS COMPARISON

Test loss measures how well a model generalizes to unseen data, with lower values indicating better predictive performance. A lower test loss suggests that the model successfully identifies meaningful patterns in salary data, while a higher test loss may indicate either underfitting (failing to capture relationships) or overfitting (learning noise instead of general trends).

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

The test losses for the three models demonstrate marginal differences. Model 1, with the simplest structure, obtained a slightly higher test loss than Model 2, indicating that its simplistic structure may not have fully captured the relationship. Model 2 achieved the best performance with a test loss of 0.1415. This result highlights its capacity to model non-linear patterns efficiently without over-complication. Meanwhile, Model 3, with two hidden layers, reported a slightly higher test loss of 0.145. Despite its increased complexity, the added parameters introduced overfitting, limiting its ability to generalize effectively to unseen data. Table 4 summarizes the test loss values for all models.

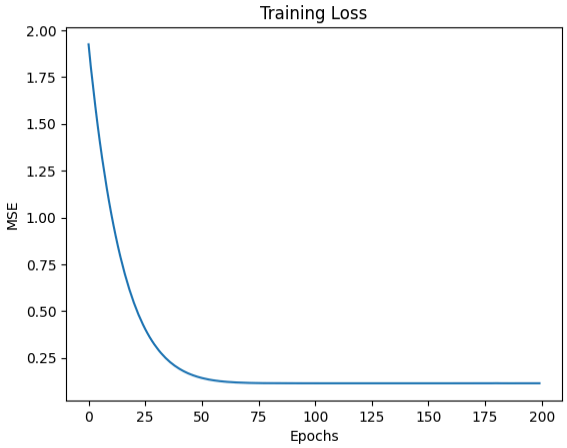
Table 4
Summary of test loss values

Model #	Test Loss Value
1	0.1435
2	0.1415
3	0.145

LOSS VERSUS EPOCHS

The training dynamics observed in the loss vs. epochs plots further illustrate the behavior of each model. Model 1 (see Figure 4) shows a smooth and gradual decrease in loss, converging to a constant value after about 60 epochs. This behavior reflects its simplicity and stable learning process.

Figure 4
Training loss graph for Model 1



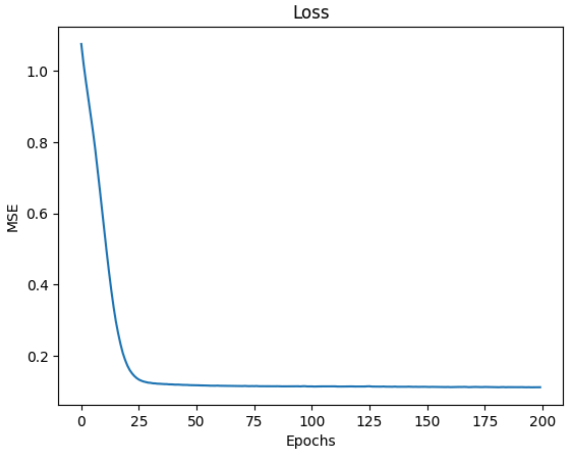
SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

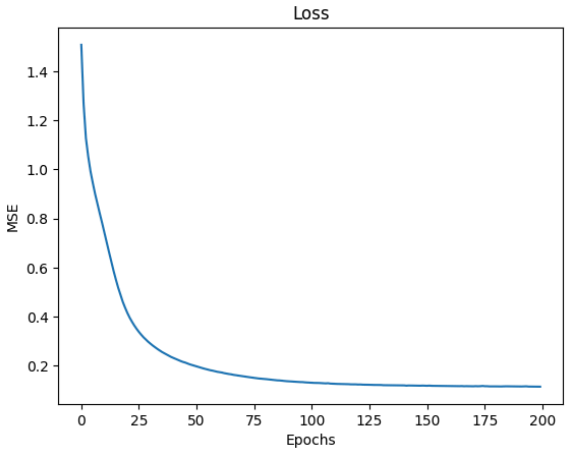
Model 2 (see Figure 5) exhibits a more dramatic initial loss reduction, stabilizing quickly around 25 epochs, which aligns with its efficient optimization and moderate complexity.

Figure 5
Training loss graph for Model 2



In contrast, Model 3's loss curve (see Figure 6) shows irregular fluctuations in the curve, converging only after approximately 100 epochs. This slower and less predictable learning behavior suggests challenges in optimizing a deeper network with more parameters and limited data size.

Figure 6
Training loss graph for Model 3



SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

The combined evaluation of test losses and visual analyses shows that Model 2 offers the most balanced performance. Its design captures essential non-linear patterns without overcomplicating the optimization process, unlike Model 3, which demonstrates signs of overfitting.

These analyses give insight into how the architecture and training dynamics of a model can influence its real-world performance. Models like Model 2, which provide both stable and accurate results, are ideal for salary prediction, ensuring consistent and dependable predictions.

OVERFITTING

Overfitting occurs when a model performs well on the training data but struggles to generalize to unseen data, often due to excessive complexity in its structure. This is evident in Model 3, where the slower convergence and marginally higher test loss suggest that the added layers and parameters did not translate into improved performance on the testing set. Instead, the model may have learned noise patterns specific to the training data, reducing its predictive accuracy on new data. In salary prediction, overfitting can lead to misleading estimates, as the model may capture noise rather than true salary trends. This reduces the reliability of the prediction and makes it lose credibility.

GENERALIZATION AND PRACTICALITY

All three models achieved reasonably low test losses, highlighting the inherent predictive strength of neural networks. Given the synthetic nature of the dataset, where patterns may be oversimplified compared to real-world scenarios, the models demonstrated an impressive ability to generalize. This simplification of patterns likely contributes to the success of the models, as it provides a more structured framework for learning.

Nonetheless, the results remain a testament to neural networks' efficiency in automating prediction tasks. Unlike manual methods, which are labor-intensive and prone to human error, neural networks leverage computational power to identify underlying relationships. Even with the limitations posed by synthetic data, this study highlights the practicality of neural networks as a superior alternative to manual prediction.

FEATURE IMPORTANCE

Permutation importance was used to evaluate the relative importance of each feature in predicting salary and their contribution to the predictions for Model 2. Given the synthetic nature of the dataset, the results are not accurate reflections of real-world data. Nonetheless, the permutation importance scores provide useful insights.

As shown in Figure 7, among the features, Education_High School emerged as the most important, with an importance score of 0.4318, followed by Education_PhD at 0.3122 and Experience at 0.1907. These results suggest that education level and experience play a significant role in predicting

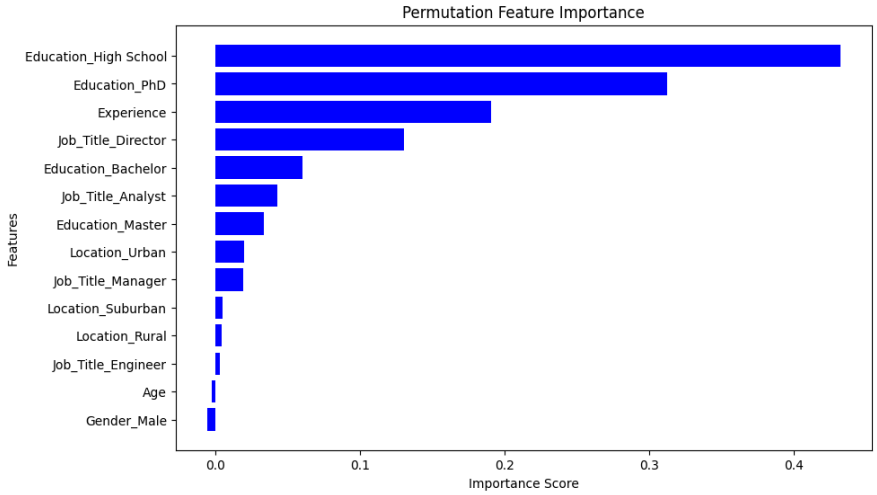
SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

salary. On the other hand, features such as Location and Gender had relatively low importance scores, indicating minimal impact on the model's predictions in this context. Age even showed a slight negative importance score of -0.0027, showing that age introduced noise in the prediction and is nearly irrelevant.

Figure 7
Permutation importance graph of Model 2



Permutation importance provides transparency in salary prediction models by identifying which factors contribute most to predictions. In real-world applications, this technique can help job seekers and employers understand the key drivers of salary decisions. For instance, if education and experience are consistently ranked as the most influential factors, individuals can make more informed career decisions by investing in relevant skills or qualifications.

CONCLUSION

This study demonstrates the capability of neural network models to predict salaries based on individual and demographic factors, using a synthetic dataset. The results indicate that a model with a single hidden layer provides the most balanced performance, effectively capturing non-linear relationships without overfitting. Despite the synthetic nature of the dataset, which may simplify real-world patterns, the models showcase the power of machine learning in automating predictions with greater efficiency compared to manual methods. This study highlights the potential of neural networks to assist in real-life decision-making, even with limitations such as data oversimplification. If applied to real-world data, the results may differ due to greater noise, missing values, and imbalanced feature distributions. Real-world data often includes outliers and has more complex, less predictable relationships influenced by socioeconomic and market-driven factors. As a result, the model's performance may decline, with lower accuracy and less clear feature importance rankings. Certain features that were unimportant in the synthetic data, like gender or location, may weigh more in a

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by ***Leadership & Innovation Lab***

SUMMER 2025

real-world context. Future research could explore more complex datasets and refine the models to handle real-world data.

REFERENCES

McKinney, K. L., Abowd, J. M., & Janicki, H. P. (2022). U.S. long-term earnings outcomes by sex, race, ethnicity, and place of birth. Quantitative Economics, 13(4), 1879–1945. https://doi.org/10.3982/qe1908

Salary prediction data. (n.d.). Kaggle. https://www.kaggle.com/datasets/mrsimple07/salary-prediction-data