Image credit to OpenAI, generated with ChatGPT.

# Does age really matter?

Programming salaries are not age dependent anymore.

## Introduction

Is age still one of the biggest predicters of salary? Or is the current digital world changing, and are we rewarded on skills and productivity instead of age? In this blogpost we will touch upon the subject of age and salary for programmers.

In the remainder of this blogpost we address the following topics:

- The dataset that was used.
- The model used for predictions.
- Model evaluation.
- Conclusions and future work.

## Dataset

The dataset used here, is the StackOverflow user survey of 2023. It contains a dataset with almost 90.000 respondents. From these 90.000 respondents. Besides salary they shared important demographic information such as:

- Age group;
- Nationality;
- Primary programming language;
- Education level;
- Years of experience.

Because we are only interested in the relation between age and salary, we cleaned up the data resulting in a data set with age groups and salaries. The following age groups were specified:

- Under 18 years of age;
- 18-24 years of age;
- 25-34 years of age;
- 35-44 years of age;
- 45-54 years of age;
- 55-64 years of age;
- 65 years of age and older.

To clean up the data we removed respondents that did not share salary information and we removed outliers that had a salary more than 3 standard deviations away from the mean. Figure 1, shows the data set. It clearly indicates that there are many outliers in this dataset.

With this dataset, we trained a machine learning algorithm that can predict salary based on the age group a person belongs to.

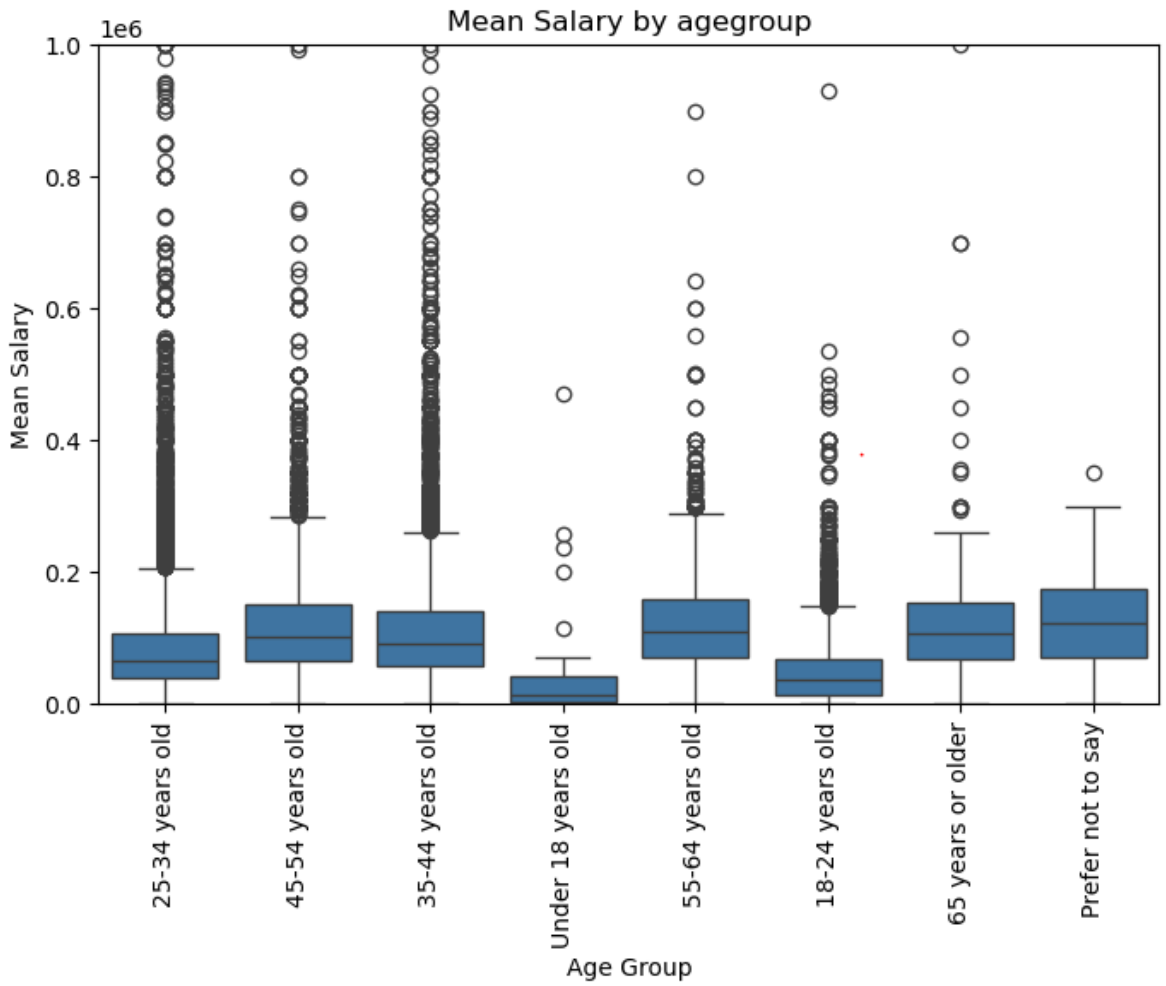


*Figure 1: Data shown in boxplots, showing the salaries per age group.*

## Performance evaluation

The trained model predicted the salary of the training set and data set with a root mean squared error (RMSE) of 93271 and 102430, respectively. From this it can be concluded that the model does slightly overfit, and is bad at predicting the salaries based on the given data. This is likely due to the fact that the age groups alone are not enough information to predict salaries accurately. Besides that it might be that age is not a good predictor of salary for programming jobs anymore.

Finally, Figure 2 shows the predicted verses the actual salaries of the test data set. This again clearly shows that the prediction quality is low. Furthermore, it shows that based on this data, the

model is very limited in making good distinctions between the different datapoints.

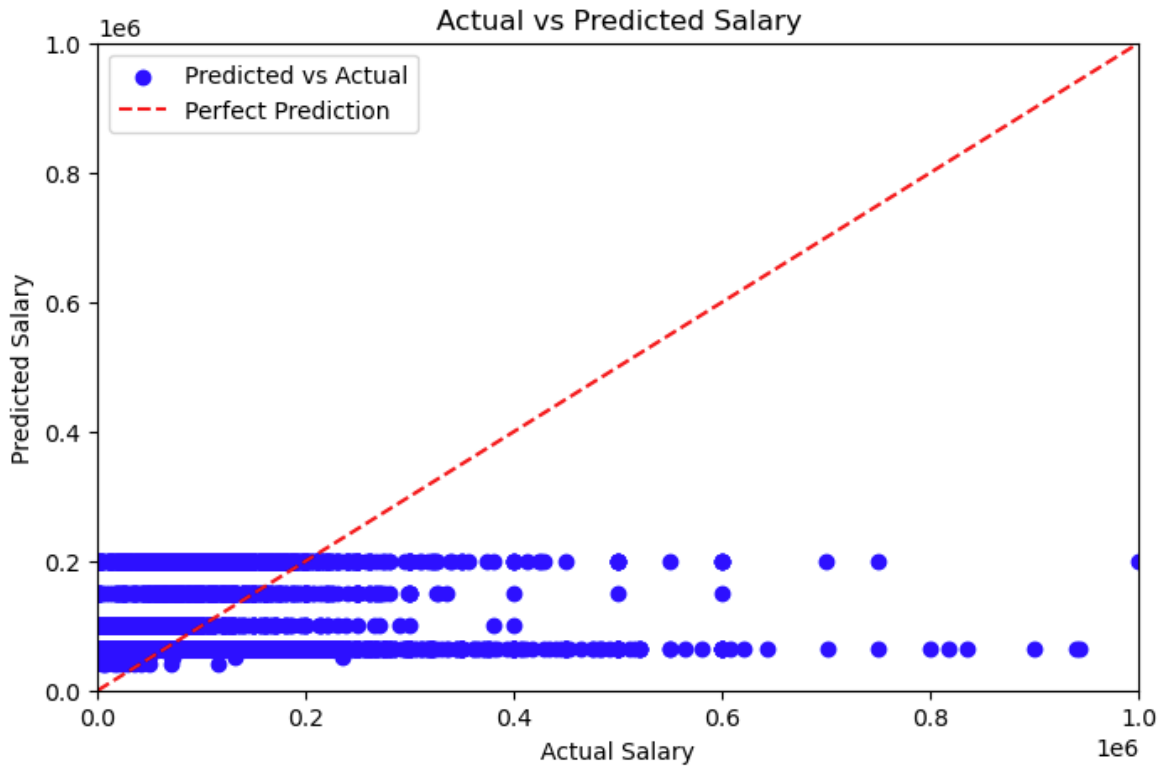


*Figure 2: predicted vs actual salaries of the test data set.*

## Conclusions and future expansions

The results presented in the blogpost show that based on age group it is not possible to accurately predict salaries of programmers. This show that age alone is not a great predictor of salaries.

For future expansions we will add more input data, such as education level, experience, and nationality to the model. This is expected to improve the model's ability to predict salary. Furthermore, this will help us to better understand which skill to improve to maximize earning potential for (new) programmers.

**Data/Scripts:** related dataset and scripts can be found in the following github repository: [JoeyRS23/UdacityCourseIntroductionToML: ProjectIntroductionToML]