**C964 Capstone: Salary Prediction Through Linear Regression**

Western Governors University

Kyle Christian - 000959459

C964 Capstone

Dr. Jim Ashe

May 21, 2023

C964: Computer Science Capstone

Table of Contents

Letter of Transmittal……………………………………………………………………………………………………………………………4

[Project Recommendation 5](#_Toc135543922)

[Problem Summary 5](#_Toc135543923)

[Application Benefits 5](#_Toc135543924)

[Application Description 5](#_Toc135543925)

[Data Description 6](#_Toc135543926)

[Objectives and Hypothesis 7](#_Toc135543927)

[Methodology 7](#_Toc135543930)

[Funding Requirements 8](#_Toc135543931)

[Data Precautions 8](#_Toc135543932)

[Developer’s Expertise 9](#_Toc135543933)

[Part B: Project Proposal 9](#_Toc135543934)

[Problem Statement 9](#_Toc135543935)

[Customer Summary 10](#_Toc135543936)

[Existing System Analysis 10](#_Toc135543937)

[Data 11](#_Toc135543938)

[Project Methodology 12](#_Toc135543939)

[Project Outcomes 13](#_Toc135543940)

[Implementation Plan 13](#_Toc135543941)

[Evaluation Plan 14](#_Toc135543942)

[Resources and Costs 15](#_Toc135543943)

[Timeline and Milestones 15](#_Toc135543944)

[Part D: Post-implementation Report 17](#_Toc135543945)

[A Business (or Organization) Vision 17](#_Toc135543946)

[Datasets 17](#_Toc135543947)

[Data Product Code 18](#_Toc135543948)

[Objective (or Hypothesis) Verification 19](#_Toc135543949)

[Effective Visualization and Reporting 19](#_Toc135543950)

[Accuracy Analysis 20](#_Toc135543951)

[Application Testing 20](#_Toc135543952)

[Application Files 21](#_Toc135543953)

[User Guide 21](#_Toc135543954)

[Summation of Learning Experience 27](#_Toc135543955)

**Letter of Transmittal**

May 13, 2023

Frodo Baggins, CTO

Middle-Earth Financial

350 Bag End Rd

Hobbiton, West Midlands

Dear Mr. Baggins,

As of recently, our competitors have developed a way to provide their customers with salary prediction, which is something we do not currently offer. Our lack of such a tool in this competitive market has caused us to lose revenue in the form of customer subscriptions. More and more customers are flocking to our competitor, Sauron Industries. We have a duty to this company and our clients to provide the most accurate and versatile system in career planning and development.

It is my strong belief that we can provide our clients with an equal or better product. This application would be a tool that our company would provide our clients that would be very easy to use and very efficient. Ideally, the client would be able to enter personal factors such as experience level in the industry and the job title they have currently or seek, and they would display a predicted salary based on similar jobs with similar experience. On the back end, this would require a nice-sized data set with machine learning to keep the recommendations as accurate as they can be.

This solution would provide our customers with an answer to one of the most important questions when it comes to choosing a career path; How much money will I make? It will benefit us as a company because it will give us more utility and be another service we can provide our clients instead of them going to someone else. This would lead to increased revenue and an increase in subscriptions.

I estimate that the project will cost around $24,000 and I would like to be the developer. I have recently finished a Bachelor’s in Computer Science and have successfully launched data tools like this one, three times prior. I believe I am more than qualified and the plan to solve our problem would be solved in a short time frame with a cost that would easily be covered in the first year of launching our new tool.

           I believe this is our best option to expand our company’s reach and capabilities regarding the need for our customers to be able to predict the salaries in their careers to come. If there are any questions regarding this, please feel free to contact me.

Sincerely,

           Kyle Christian

## Project Recommendation

Problem Summary

Middle-Earth Financial has tried to always be the cutting edge when it comes to finding its clients’ career paths that will be fruitful based on factors such as experience and the job title they wish to end up with. Recently, its competitor Sauron Industries has come out with a tool that provides its clients the ability to estimate their salary which has led to us losing some of our customers due to our lack of such a tool.

Application Benefits

The primary benefit of the proposed application would be our ability to stay competitive in an ever-evolving market. The new data product will give us the ability to provide a service that is invaluable in choosing a career. The application will drive more customers to our company and allow us to retain more of the customers we have currently. The value of our company as a whole will increase and the increased revenue will benefit the stakeholders.

Application Description

The application will be a cloud-computed notebook hosted by Kaggle.com. The programming language of choice will be Python as it is currently the leading language for data science. The interface will feature the title “Salary Prediction” and have a drop-down menu featuring different levels of experience. The user will choose a selection that most closely resembles their experience in the cyber security sector. The application will use linear regression and provide the user with a salary prediction based on said experience while also providing visual aids for the user.

Data Description

The data that will be used in the solution will be provided by Kaggle.com. The dataset is a cyber security dataset that provides job data such as:

* Work Year (Integer)
* Experience Level (String)
* Employment Type (String)
* Job Title (String)
* Salary (Integer)
* Salary Currency (String)
* Salary in USD (Integer)
* Employee Residence (String)
* Remote Ratio (Integer)
* Company Location (String)
* Company Size (String)

Objectives and Hypothesis

There are multiple objectives for this new data product. The most important objective is to have a working tool that can semi-accurately predict a salary based on user input. A secondary objective for this project is to meet all the deadlines put in place. Another objective is to make sure that there are enough visuals in place so that the user understands the correlation between experience level and salary within the field of cybersecurity.

I hypothesize that by using Linear Regression, this application will be able to achieve semi-accurate salary predictions based on users inputting their experience level. This product will help our company in numerous ways and will return the cost of the product within less than a year.

Methodology

The development will follow the SEMMA methodology.

• Sample: A large dataset from Kaggle.com will be extracted and a sample that represents all the data will be taken out. This is where we will select the variables and divide the data into training and validation subsets. This step will reduce the cost and time taken to process.

• Explore: The salary data will be explored for outliers and anomalies to better understand the data. The data will be plotted on a graph to find the correlation between salaries and the variables that were selected during the sample phase.

• Modify: In this step, we will be correcting errors and inconsistencies or restructuring data to make it easier to use aka “cleaning” the variables mentioned above. This will make sure that the model in the next step is more and more accurate in each iteration.

• Model: This project will aim to use a linear regression model. A linear regression model will provide the ability to predict salary based on the factors that were chosen in the sample phase and modified in the modification phase.

• Assess: Compare data to the test data for accuracy. Find solutions to overcome the limitations of the program.

Funding Requirements

The funding for this data project will be simple due to its low requirements. There will be one developer (me) and the salary will amount to $15,000. This project is estimated to take 150 hours and at the hourly rate of $100, it amounts to $15,000. In addition to the cost of the developer, there is the need for a new workstation that will amount to $4,000. Our current workstations are out of date and not equipped for the project. Lastly, the tool will need to be maintained over the first year, which will cost $5,000. At $50 an hour to maintain 100 hours in the first year it amounts to $5,000. There will be no funding requirements for the data because we will be using a public dataset and public libraries.

Data Precautions

There will be no data privacy issues since the data will be acquired from a public data source, Kaggle.com. Kaggle requires that the data from their website not be used commercially, and we will not be using their data in this manner. Compliance with their terms also states that we will not be using any data to create obligations under HIPPA (Health Insurance Portability and Accountability Act.)

Developer’s Expertise

I have chosen to personally handle the development of this new data product. I have been honing my skills in software development in and outside of multiple projects during my time at Western Governor’s University. The degree I acquired from Western Governors was a Bachelor of Computer Science. With the knowledge I attained from the degree and the work I have done solo, I believe I would be a great fit for developing this new data product.

# Part B: Project Proposal

## Problem Statement

As of recently, our competitors have developed a way to provide their customers with salary prediction, which is something we do not currently offer. Our lack of such a tool in this competitive market has caused us to lose revenue in the form of customer subscriptions. More and more customers are flocking to our competitor, Sauron Industries. We have a duty to this company and our clients to provide the most accurate and versatile system in career planning and development.

This project proposes to use already established data regarding the salaries of infosec professionals as well as the ability to gather new data to help better predict the salaries based on experience level, job title, or years they have worked. The new data will help train the algorithm and minimize human interaction.

## Customer Summary

The customers that will be impacted by the proposed data product are our current and future clients who use the Career Planning tools we offer at Middle-Earth Financial. The clients are looking to either start a new and lucrative career or make decisions that can get them to move upward at their current place of employment. Our clients range in age, from 18 to senior citizens. This means that there is a large variety of experience levels including entry-level, mid-level, experienced, and senior-level. All our customers are interested in or are a part of, the cybersecurity sector, so the tool will relate only to cybersecurity.

The new data product will be able to cater to the different aspects mentioned above. New and old clients alike will be able to access the new data product in the usual sections they visit. There will be a link to the tool that they can click on and use immediately. The tool will have a user-friendly interface that allows the client to enter their experience level in cybersecurity and will return a salary to give the user an idea of what they should be making or can expect to make in such a position, with that level of experience. This will help guide our clients to a better career and a better life while also increasing the value of our company by attracting new clients and gaining more and more subscriptions.

## Existing System Analysis

Our company currently offers various tools for client to help them get connected with job opportunities in the cyber security sector. Our job search tool allows our clients to find jobs in cyber security that have been posted by an assortment of tech companies. At the same time, our employer section allows employers to add jobs that they need to be filled by adding a title, requirements, and recommended credentials. The clients that use the employer section are also able to filter through applicants with filters like “Only applicants with photos” and 3-5 years of experience. The clients that are applicants can filter jobs based on different criteria they deem to be important. Some examples of the filters are “Urgently hiring” and “Seeking entry-level experience.”

Another tool our company currently offers is a path tool in which the user can pick a job title and the tool will show them a list of steps that will likely result in them achieving that path. The path explains each step in detail and describes the likely duties they will have if they achieve the job they desire. An example of this is if the user chooses “Penetration Tester” as their end goal, the program will tell them first they need to acquire knowledge of networks, hacking tools, etc. It will also lay out a job order in which they might make it to their end goal. Ex: Network Administrator to Cyber Security Analyst to Penetration Tester.

## Data

Our new data product will be using a dataset from the public data source Kaggle.com. Kaggle is an online community for data science and machine learning (ML) enthusiasts. Owned by Google, it is currently the world’s largest crowdsourced web platform for data scientists and ML practitioners. Thus, Kaggle gives you access to several professionals in your field whom you can brainstorm, compete, and solve real-life problems with. (Adegoke, 2023, para. 3)

The dataset that we will use has already been cleaned of Null values and contains many different variables. The data will be pulled in and stored on our local machine by using Kaggle’s API to download and save the table. The table includes the work year, experience level, employment type, job title, salary, salary currency, salary in USD, employee residence, remote ratio, company location, and company size. Work year, salary, salary in USD, and remote ratio are integer values while the rest are string data. This is a lot of data, and we will need to filter out certain values to increase accuracy and focus our product on the experience level using various methods in the Sklearn library of Python.

## Project Methodology

The development will follow the SEMMA methodology.

• Sample: A large dataset from Kaggle.com will be extracted and a sample that represents all the data will be taken out. This is where we will select the variables and divide the data into training and validation subsets. This step will reduce the cost and time taken to process.

• Explore: The salary data will be explored for outliers and anomalies to better understand the data. The data will be plotted on a graph to find the correlation between salaries and the variables that were selected during the sample phase.

• Modify: In this step, we will be correcting errors and inconsistencies or restructuring data to make it easier to use aka “cleaning” the variables mentioned above. This will make sure that the model in the next step is more and more accurate in each iteration.

• Model: This project will aim to use a linear regression model. A linear regression model will provide the ability to predict salary based on the factors that were chosen in the sample phase and modified in the modification phase.

• Assess: Compare data to the test data for accuracy. Find solutions to overcome the limitations of the program.

## Project Outcomes

The project will have only a few deliverables as it is not very large. One of those deliverables will be the data product itself which will need to be working. After the data product is working, it will need to be tuned for accuracy as we want our clients to have semi-accurate salary predictions. After the product is complete, we will need a user guide that explains step-by-step how to use the new data product. There will need to be documentation on the data product that explains each aspect of the core functionality for our programmers that will later be maintaining and updating the code.

## Implementation Plan

The plan for integrating this new data product into our existing systems upon completion is set below.

**Test Environment** - Our new tool will be released on a small set of computers which will be our release test environment. Upon release on the test environment, end users will test the tool and make sure that it is working properly and report any issues with the current tool or other systems.

**Employee Training** – After the test environment is successful, our employees will need to be trained in the essentials of the new application. Training will be held to go over the essential functions of the software, how to use it, and who should be contacted in case of software failure.

**Rollout** – This will be a full release integrating the new tool into our systems for our clients to be able to access it regularly. We will be constantly assessing our systems to ensure that the release hasn’t caused any other issues with any of our other programs. A team will be dedicated to this purpose.

**Feedback** – We will gather feedback from our team and the end users as to how the release went, how the tool works for them, and what we can do better in the future when we release a new data product.

## Evaluation Plan

The first step in evaluating our product will be to form an evaluation team of our stakeholders and senior IT members. The software will then be evaluated by answering a series of questions.

**Did our new data product fix our problem?** – We will return to the problem at the origin and assess whether our new application solved our initial problem.

**Are there any other areas where it can help?** – We will ask the stakeholders if they believe the application can be of any other use to maximize the potential of the program and its use to us as a company.

**Did we meet our goals and milestones?** – An assessment of the scheduling and the completion of the milestones will take place to see where and when there was a success and where there was a failure.

**Is the application semi-accurate?** – We will assess the R2 scoring of the application and see that it meets a good score for our customers and if it doesn’t we will need to increase our accuracy.

## Resources and Costs

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Workstation | Computer | 4,000 |
| Lead Developer | Programmer | 15,000 |
| Maintenance | First Year | 5,000 |
|  | **Total** | 24,000 |

.

## Timeline and Milestones

The project will take roughly two months from beginning to finish. The plan is to have this project completed by May 23, 2023. The project is estimated to take around 150 hours which is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone | Start and End Dates | Duration | Resources |
| Analyze Requirements | 05/01/2023-05/07/2023 | 10 hours | Stakeholders |
| Code Development | 05/08/2023-05/10/2023 | 10 hours | Software Dev |
| Gui Development | 05/11/2023-05/12/2023 | 8 hours | Software Dev |
| Application Testing | 05/13/2023-05/18/2023 | 40 hours | Software Dev / QA |
| Documentation | 05/19/2023-05/23/2023 | 40 hours | Software Dev |
| Systems Integration | 06/01/2023-06/14/2023 | 42 hours | Software Dev |

**References**

Adegoke, J. (2023, April 17). *A Beginner’s Guide to Kaggle for Data Science*. MUO. https://www.makeuseof.com/beginners-guide-to-kaggle/

Part C: Application

Being that the entire data product is in a Jupyter Notebook, it only requires one file.

\capstone.ipynb Jupyter Notebook File

It can be downloaded via <https://github.com/tweaksneak/Salary-Prediction>

It can be run using Jupyter Lab downloaded from <https://www.anaconda.com/>

# Part D: Post-implementation Report

## A Business (or Organization) Vision

Before recently, our competitors had developed a way to provide their customers with salary prediction, which is something we did not offer. Our lack of such a tool in this competitive market had caused us to lose revenue in the form of customer subscriptions. More and more customers were flocking to our competitor, Sauron Industries. We had a duty to this company and our clients to provide the most accurate and versatile system in career planning and development.

Our project used already established data regarding the salaries of infosec professionals as well as the ability to gather new data to help better predict the salaries based on experience level, job title, or years they have worked. We were able to successfully create a data product that can semi-accurately predict a salary based on experience level in the cyber security sector.

## Datasets

The dataset that we used has already been cleaned of Null values and contained many different variables. The data was pulled in and stored on our local machine by using Kaggle’s API to download and save the table. The table included the work year, experience level, employment type, job title, salary, salary currency, salary in USD, employee residence, remote ratio, company location, and company size. Work year, salary, salary in USD, and remote ratio were integer values while the rest were string data. This was a lot of data, and we needed to filter out certain values to increase accuracy and focus our product on experience level using various methods in the Sklearn library of Python.

Being that we wanted to focus on experience level as the primary variable, we encountered an issue. Linear regression models use numerals as input and the experience level was given to us as integer data. After researching how to handle this, I chose to use the One Hot Encoding method which extends the columns of the table by the number of unique choices. This turned the four options of experience “entry-level”, “middle level”, “experienced”, and “senior level” into four separate columns giving them a value of 0 or 1. 0 would be the value in the table if the row in the table did not have that level of experience and 1 would be the value in the table if the user had that experience.

## Data Product Code

The data product code first took a set of raw data that contains 11 different variables. Most of these variables were string data which was then filtered out. The goal was to use linear regression which requires that the values be numerals. Unfortunately, the data for the experience level (which was to be the independent variable) was in string format. I used the one hot encoding technique which created four extra columns in the data set. Values of 1 or 0 were given to the new columns based on whether the row contained that level of experience or not.

The descriptive methods and visualizations I chose were two different bar plots and a pie chart. All three of these visualizations bring information to the user regarding the connection between salary and experience levels.

The non-descriptive method was extracting a salary prediction based on the data we had. For this, I used a linear regression model with an R^2 scoring method.

## Objective (or Hypothesis) Verification

I hypothesized that by using Linear Regression, this application would be able to achieve semi-accurate salary predictions based on users inputting their experience level and that this product would help our company in numerous ways and would return the cost of the product within less than a year.

Part of the hypothesis was met in the form of the application achieving semi-accurate salary predictions based on user input and a linear regression model. As far as the cost of the product being returned to the company within the first year is concerned, we have yet to see this as it has not been a year yet.

## Effective Visualization and Reporting

My descriptive methods and visualizations supported the development process in several ways. I was able to explore the data and see how it all connected. Being that I wanted to show a correlation between experience level and salary, I needed to be able to analyze the data. Being able to see it visually was imperative.

Visualization #1 was a bar plot showing how many jobs there were relating to each experience level. This gave me the ability to understand the data more and understand the relationship with experience level.

Visualization #2 was a bar plot showing the average salary for each experience level. This was important to see visually because later the prediction would be gauged against the average to get a ballpark idea of how accurate the predictions were, even though later we would use an R2 scoring model.

Visualization #3 was a pie chart that showed the percentage of jobs that were in each category of experience level. This was another way to visualize the difference between the experience levels and analyze the data.

## Accuracy Analysis

The accuracy function I used to test the linear regression model and the dataset I chose is the R^2 score from Scikit-Learn. When it comes to the R^2 score or (coefficient of determination) a score of 1.0 is the best possible score.

The function called to measure the R^2 was: r2\_score(y\_test, y\_pred)

The y\_test represented the test data and the y\_pred represented the prediction data using the linear regression model. The score achieved was .46 which is semi-accurate.

## Application Testing

The first testing that took place was making sure there were no syntax errors as the code was developed. Any errors that became present as each section of code was developed were dealt with immediately so as not to stack up multiple errors and make it more difficult to deal with them. The next testing was logical in nature to make sure that the application was accurately predicting salary based on the user’s input. This was done using the R2\_score method of Sklearn. After seeing an R2 score and realizing it was not super accurate, I attempted to fine-tune the data to achieve a higher score by eliminating certain variables or adding some in. After testing the syntax and making sure that the program did what it was intended to do, there was no more testing that was necessary.

## Application Files

Being that the entire data product is in a Jupyter Notebook, it only requires one file.

\capstone.ipynb Jupyter Notebook File

## User Guide

Step 0. Navigate to <https://github.com/tweaksneak/Salary-Prediction> and download the ZIP

A screenshot of a computer

Description automatically generated with medium confidence

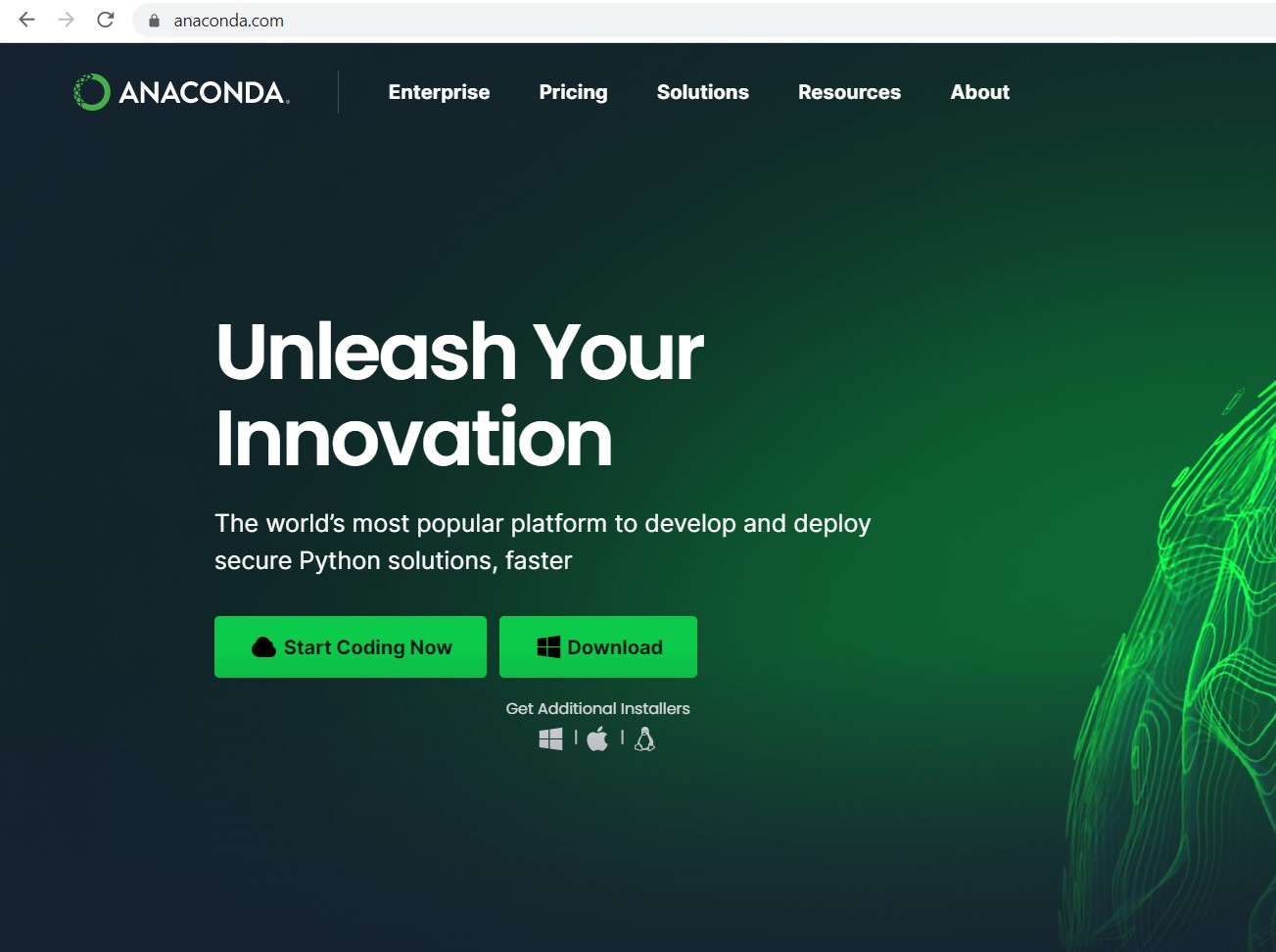
Step 1. Extract all from the newly downloaded ZIP file

A screenshot of a computer

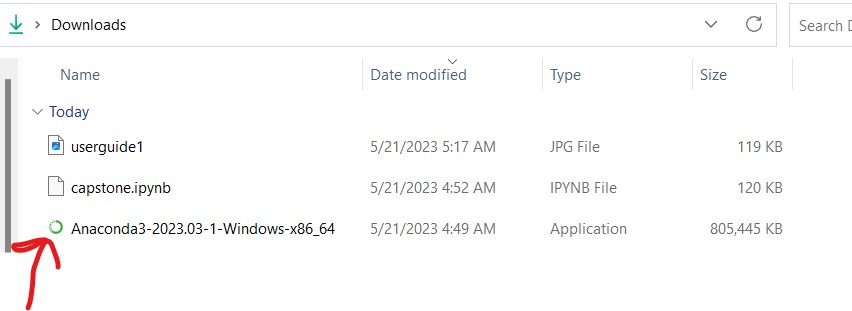
Description automatically generated

Step 2. Navigate to <https://www.anaconda.com/>

Step 3. Press download



Step 4. Navigate to the default download folder for your system and run the installation



Step 5. Follow the default installation prompts to finish installing

Step 6. Use the Windows search bar to search for Anaconda

A screenshot of a computer

Description automatically generated

Step 7. Double-click Anaconda Navigator and wait for it to load.

Step 8. Click on the launch button for Jupyter Labs

A screenshot of a computer

Description automatically generated

Step 7. Click the upload file button.

A screenshot of a computer

Description automatically generated with medium confidence

Step 8. Select the downloaded and extracted capstone.ipynb file from step 1.

Step 9. Double click capstone.ipynb in the side panel

A screenshot of a computer

Description automatically generated with medium confidence

Step 10. Click to restart kernel and run all cells

A screenshot of a computer

Description automatically generated with medium confidence

Step 11. Scroll the notebook and look at all the information. A drop down menu and button are available for you to predict salary based on your experience level!

A screenshot of a computer

Description automatically generated with medium confidence

## Summation of Learning Experience

My only experience before getting to this project was the courses I have taken here at WGU and those I took at previous education establishments. I had no prior knowledge of how to complete this project before I started my degree. I used the videos that are offered by the course instructor and the Capstone Website that was created to help students complete this project. I was able to use my knowledge from prior classes that involved Python to help me in making sure that my syntax was correct when piecing together the code portion of the project. I also took a look at the example projects to get an idea of how I should lay out my documentation. Another thing I was completely unfamiliar with was Jupyter Notebooks. I had to read the documentation and navigate different websites attempting different methodologies. I attempted to use Kaggle Notebooks and finished the data product but had issues with the ipywidgets. I then had to change to a standalone Jupyter Notebook through Anaconda. With all of these things, I was able to complete my project and finish my class.