01/20/2024

C964: Computer Science Capstone

Salary Prediction Model

Western Governor’s University

Jeff Fenwick

Student ID #: 001549011

Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 3](#_Toc157511405)

[Part B: Project Proposal Plan 5](#_Toc157511406)

[Project Summary 5](#_Toc157511407)

[Data Summary 6](#_Toc157511408)

[Implementation 7](#_Toc157511409)

[Timeline 9](#_Toc157511410)

[Evaluation Plan 9](#_Toc157511411)

[Resources and Costs 10](#_Toc157511412)

[Part C: Application 11](#_Toc157511413)

[Part D: Post-implementation Report 12](#_Toc157511414)

[Solution Summary 12](#_Toc157511415)

[Data Summary 12](#_Toc157511416)

[Machine Learning 14](#_Toc157511417)

[Validation 15](#_Toc157511418)

[Visualizations 16](#_Toc157511419)

[User Guide 17](#_Toc157511420)

[Reference Page 19](#_Toc157511421)

# Part A: Letter of Transmittal

01/22/2024

Emily Johnson

QuantumCipher Analytics Co.

123 Main Street

San Francisco, CA 94101

Dear Mrs. Johnson,

As our company grows and evolves in the dynamic field of data science, we face the critical task of establishing fair and appealing compensation packages. This issue could potentially impede our ability to attract and retain top talent, possibly losing our competitive edge in the market. Our organization needs to develop competitive compensation strategies and I have found a solution. The solution is to develop a comprehensive application that utilizes data analytics to assess market trends, based on employee demographics and skill levels. This application will enable the company to create a balanced salary structure.

The implementation of this solution will streamline our salary structure and enhance transparency and fairness in compensation across the company. It will improve employee satisfaction, reduce turnover, and position our company as a top employer in the growing industry of data science. This tool will also allow for agile responses to market changes and will help keep our compensation strategies aligned with industry standards.

The project cost estimate is $125,000, including software development, data acquisition, and testing phases. The timeline for this project is January 29th through April 11th. The team will use data from a combination of surveys and job listings with disclosed salary information which is gathered from an online source and will continue to grow and become better as more data is acquired. Regarding the data, we are committed to ensuring that the development of this application will adhere to the highest ethical standards. In addition to ensuring data privacy and security, we will also make sure that our compensation strategy is free from biases.

With several years of experience in completing internal projects and an understanding of both data analytics and HR processes at QuantumCipher Analytics Co., I am perfect to lead this initiative. I am enthusiastic about the potential impact of this proposal on our organization and am looking forward to discussing this opportunity further. Your support in this initiative is crucial and I am available to provide any additional information as needed.

Thank you for considering this proposal. I am confident that with our combined efforts, QuantumCipher Analytics Co. can set a new standard in employee satisfaction in the data science sector.

Sincerely,



Jeff Fenwick, Internal Projects Engineer

# Part B: Project Proposal Plan

## Project Summary

* **Describe the problem.**

*QuantumCipher Analytics Co. faces a significant challenge in determining competitive salaries for its rapidly growing team of data scientists. This task is crucial in attracting and retaining top talent, especially in a competitive market where numerous companies vie for skilled professionals. The main issue lies in building compensation packages that are both fair and appealing, using variables such as experience, geographic location, and the size of the company.*

* **Summarize the client and their needs as related to the problem.**

*The client itself is QuantumCipher Analytics Co. This is an internal project designed to assist the company with remaining competitive. Their need centers around effectively balancing compensation to attract and retain skilled professionals while staying financially viable in a competitive market.*

* **Provide descriptions of all deliverables. For example, the finished application and a user guide.**
  + Multiple Regression Machine Learning Model

*A developed model to predict data scientist salaries based on independent variables. This model will be implemented in an application to be used with JupyterLab. The application will take independent variables such as experience and location and determine an appropriate salary.*

* + User Guide

*Detailed and specific instructions on how to access and use the application integrated with the machine learning model.*

* + Project Documentation

*Comprehensive documentation covering all aspects of the project, including technical details and user instructions.*

* **Provide a summary justifying how the application will benefit the client.**

*The application will benefit QuantumCipher Analytics Co. by enabling precise and data-driven salary determinations for their data science team. This application ensures competitive and equitable compensation, which is essential for attracting and retaining top talent in a highly competitive market.*

## Data Summary

* **Provide the source of the raw data, how the data will be collected, or how it will be simulated.**

*The raw data for this project was sourced from Kaggle.com, specifically the "Data Science Job Salaries" dataset. This dataset is a compilation of over 600 entries, each entry detailing various aspects of data science job positions. The data encompasses a range of columns: Work Year, Experience Level, Employment Type, Job Title, Salary, Salary Currency, Salary in USD, Employee Residence, Remote Ratio, Company Location, and Company Size. The original data was found on https://ai-jobs.net/salaries/, gathered through a combination of surveys and job listings with disclosed salary information. This comprehensive dataset provides a broad view of the data science job market, particularly focusing on salary variations across different factors such as experience, location, and company size.*

* **Describe how data will be processed and managed throughout the application development life cycle: design, development, maintenance, or others.**

*In the preprocessing stage, the dataset will undergo several modifications to align with the project’s specific needs and to ensure data quality. Redundant columns will be removed, as well as the “Job Title” column which can skew the analysis.*

*To maintain relevance to the project’s business case, which is centered on a U.S. company, the data will be filtered to include only those entries where the company location is in the United States. This filtering helps in focusing the analysis on the geographical context of interest. Outliers will be filtered based on groups in each column. The extreme salaries will be removed, which will increase the accuracy.*

*Additionally, an unnamed index column present in the dataset will also be removed. This column, typically used for identification purposes, will not be necessary for our analysis and its removal will help in simplifying the dataset.*

*Over the lifecycle of the application, more data will be used to increase effectiveness, and this may lead to more categories and columns of data. This can also lead to previously unused portions of data being used (with more part-time data, that option can be re-admitted). Each new potential data category will need to be reviewed thoroughly to be accepted. This will happen during development and maintenance.*

* **Justify why the data meets the needs of the project. If relevant, describe how data anomalies, e.g., outliers, incomplete data, etc., will be handled.**

*The data meets the needs of the project because it provides comprehensive and relevant information for making data-driven salary determinations. It enables (relatively) precise salary decisions for the data science team which is essential for remaining competitive and attracting and retaining top talent.*

*Data anomalies will happen in salary data because there are many factors, some of them subjective, that influence a person’s salary. Outliers must be handled correctly to get a better sense of the industry. The data anomalies in this set of data are found by looking for extreme highs and lows through each category. For example, looking at all the small business, the extreme high and low salaries from all data entries for small business sizes will be removed from the cleaned dataset. This goes a step beyond just eliminating all high and low salary data from the dataset.*

*There was no incomplete data in the dataset. When more data is collected, it will have to be analyzed to ensure that missing and incomplete data entries are handled correctly.*

* **Address any ethical or legal concerns regarding the data. If there are no concerns, explain why.**

*There are no ethical or legal concerns regarding the data because there is no sensitive or proprietary data used. There is no personal identifying information and even if there was, it wouldn’t be relevant so it wouldn’t be used.*

## Implementation

* **Describe an industry-standard methodology to be used.**

*The methodology that will be used to implement this project will be the SEMMA methodology. This is a process designed to create effective predictive and descriptive models. First, a sample is gathered. This sample is then explored to identify the data’s characteristics and identify anomalies. Next, the data is modified, or preprocessed, which prepares it for modeling. Then, the data is fit to the model using machine learning techniques. Finally, the model is assessed, and its effectiveness is evaluated.*

* **An outline of the project’s implementation plan. This outline can focus on the project’s development as a whole; or it may focus on only the implementation of the machine learning solution.**

*This outline will focus on the implementation of the machine learning solution:*

* *Sample*

*A subset of the data will be selected to analyzed. The rest will be used for testing.*

* *Explore*

*The data will be visualized, and the descriptive statistics will be calculated to identify outliers and anomalies.*

* *Modify*

*The data will be preprocessed, or cleaned, to prepare it for modeling. This will remove outliers and categories unfit for use in the model. Dummy variables will also be created and used for categorical data.*

* *Model*

*The predictive model will be trained on the data selected for analysis. This model will be created using multiple regression and will predict a dependent variable (salary in USD) given multiple independent variables.*

* *Assess*

*The model’s effectiveness will be evaluated using the test data. The accuracy will be determined by the Mean Squares Error and the R-squared value. Because of the subjective nature of salary values, less accuracy is required for the model to be effective. The main win condition of this model is to develop a working foundation to add more data to.*

## Timeline

* **Provide a projected timeline, including projected start dates and end dates for each milestone (a table is not required but encouraged).**

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Project Initiation | 2 days | 01/29/2024 | 01/31/2024 |
| Data Collection | 5 days | 02/01/2024 | 02/06/2024 |
| Data Exploration | 7 days | 02/07/2024 | 02/14/2024 |
| Data Preprocessing | 10 days | 02/15/2024 | 02/25/2024 |
| Model Development | 14 days | 02/26/2024 | 03/11/2024 |
| Model Validation | 7 days | 03/12/2024 | 03/19/2024 |
| Assessment and Refinement | 10 days | 03/20/2024 | 03/30/2024 |
| Finalization and Documentation | 5 days | 03/31/2024 | 04/05/2024 |
| Deployment | 5 days | 04/06/2024 | 04/11/2024 |

## Evaluation Plan

* **Describe the verification method(s) to be used at each stage of development.**
  + *Sample*

*Check for missing values or inconsistent values.*

* + *Explore*

*This is tied into the sample phase. This can also involve visualizations to ensure the data is consistent to real-world experience.*

* + *Modify*

*Check for successful handling of the outliers by verifying that the number of data entries has changed. Use a known outlier to verify that the modifications are correct.*

* + *Model*

*Using the application tool, verify that the model works by inputting variables and having the model predict the salary. Verify that the prediction is aligned with real-world expectations.*

* + *Assess*

*Use the test data to validate accuracy of the model. Use different randomize seeds to check consistency across the dataset.*

* **Describe the validation method to be used upon completion of the project.**

*Upon completion of the project, the validation method that will be used to gauge the project’s success will be a performance evaluation. This evaluation will use the test set of data and will assess the model’s performance using Mean Squared Error and R-squared metrics. Successful completion will not rely solely on accuracy as the purpose of this model is to be a foundation for more data. If the model can generate predictions aligned with real-world expectations consistently, the project will be ready for the next phase, which will be to acquire more data.*

## Resources and Costs

* **Itemize hardware and software costs.**

|  |  |
| --- | --- |
| Workstation for Data Analysis | $1,500 - $2,000 |
| Data Storage | $100 - $300 |
| Data Analysis and Modeling Software | $100 - $1,000 |
| Version Control Software | $100 - $250 |
| Total | $1,950 - $3,550 |

* **Itemize estimated labor time and costs.**

|  |  |
| --- | --- |
| Data Scientist | $40,000 - $60,000 |
| Software Developer | $30,000 - $45,000 |
| Project Manager | $16,000 - $24,000 |
| QA Engineer | $14,000 - $20,000 |
| Total | $100,000 - $149,000 |

* **Itemize estimated environment costs of the application, e.g., deployment, hosting, maintenance, etc.**

|  |  |
| --- | --- |
| Cloud Computing and Storage | $100 - $1,000 |
| Development and Testing Environments | $50 - $300 |
| Backup and Disaster Recovery Solutions | $30 - $300 |
| Security and Compliance | $20 - $200 |
| Total | $200 - $1,800 |

# Part C: Application

**Submitted Files**

* **C964\_Capstone\_Project.ipynb**

*I am including the local version of the JupyterLab notebook file as well in case the JupyterLite version doesn’t work. This one requires additional libraries but works fine with the Anaconda Distribution of Python found* [*here*](https://www.anaconda.com/download)*. Once this is installed, search for Anaconda Navigator and launch it. Open JupyterLab, then load this file as well as the csv file. Select Run -> Restart Kernel and Run All Cells.*

* **C964\_Capstone\_Project\_JupyterLite.ipynb**

*This should be uploaded to the JupyterLite online platform along with the csv file per the user manual below.*

* **ds\_salaries.csv**

*This data file from Kaggle.com contains the data used for the training and test sets. This should be uploaded to JupyterLab or JupyterLite so that it can be read by pandas.*

# Part D: Post-implementation Report

## Solution Summary

* **Summarize the problem and solution.**

*The problem stems from a challenge faced by the company, QuantumCipher Analytics Co. The company needed to determine competitive salaries for their rapidly expanding team of data scientists. Because this is such a large company with many competitors, the company recognized the necessity of offering competitive compensation to attract and retain top talent. To tackle this issue, the company turned to a data-driven approach, utilizing a comprehensive dataset of data science job salaries. This dataset provided the data necessary for developing a robust solution.*

*The solution to this problem was the development and application of a multiple regression machine learning model. This was completed and is operational and provides prediction of data scientist salaries based on a variety of independent variables, such as experience level, location, and company size.*

* **Describe how the application provides a solution to the problem from parts A and B.**

*The application allows the company to provide competitive and fair compensation to their employees by aiding in the strategic decision-making regarding employee salaries and enhancing the company’s ability to maintain its position as a leader in the industry. The model’s predictive capabilities, grounded in rigorous data analysis, position the company to effectively navigate the intricacies of salary determination in the field of data science.*

## Data Summary

* **Provide the source of the raw data, how the data was collected, or how it was simulated.**

*The raw data for this project was sourced from Kaggle.com, specifically the "Data Science Job Salaries" dataset. This dataset is a compilation of over 600 entries, each entry detailing various aspects of data science job positions. The data encompasses a range of columns: Work Year, Experience Level, Employment Type, Job Title, Salary, Salary Currency, Salary in USD, Employee Residence, Remote Ratio, Company Location, and Company Size. The original data was found on https://ai-jobs.net/salaries/, gathered through a combination of surveys and job listings with disclosed salary information. This comprehensive dataset provides a broad view of the data science job market, particularly focusing on salary variations across different factors such as experience, location, and company size.*

* **Describe how data was processed and managed throughout the application development life cycle: design, development, maintenance, or others.**

*In the preprocessing stage, the dataset underwent several modifications to align with the project’s specific needs and to ensure data quality. Firstly, the 'Job Title' column was removed. This decision was based on the assessment that job titles, in this context, do not significantly correlate with salary variations and could introduce unnecessary complexity into the analysis. There are some titles that do convey an increased salary, such as those containing “Director”, but most do not.*

*To maintain relevance to the project’s business case, which is centered on a U.S. company, the data was filtered to include only those entries where the company location is in the United States. This filtering helps in focusing the analysis on the geographical context of interest. Additionally, data points corresponding to part-time employment were excluded due to their insufficient representation in the dataset. Finally, a critical step in the preprocessing was filtering out outliers. These outliers were filtered based on groups in each column. The extreme salaries were removed and the accuracy increased.*

*To streamline the data and avoid redundancy, the 'Salary' and 'Salary Currency' columns were removed, as these were deemed superfluous given the presence of a 'Salary in USD' column which provided a standardized measure of salaries across different geographies. This standardization is crucial for consistent and comparable analysis, especially when dealing with a global dataset.*

*Additionally, an unnamed index column present in the dataset was also removed. This column, typically used for identification purposes, was not necessary for our analysis and its removal helped in simplifying the dataset. These steps were instrumental in ensuring that the dataset was not only relevant to the specific context of the study but also free from unnecessary complexities that could potentially skew the analysis or interpretation of the results.*

*Over the lifecycle of the application, more data will be used to increase effectiveness, and this may lead to more categories and columns of data. This can also lead to previously unused portions of data being used (with more part-time data, that option can be re-admitted). Each new potential data category will need to be reviewed thoroughly to be accepted. This will happen during development and maintenance.*

## Machine Learning

* **Identify the method and what it does (the “what”).**

*The method used for machine learning on this project was a multiple regression model. This was implemented through sklearn.linear\_model import LinearRegression. This method is used to understand the relationship between one dependent variable (salary in USD) and two or more independent variables. The value of the dependent variable is predicted based on the combined influence of the independent variables.*

* **Describe how the method was developed (the “how”).**

*The model calculates the coefficients for each independent variable, providing a number of their individual contribution to the dependent variable. These coefficients are derived from using training data and fitting the model to the data using the least squares method, which “minimizes the sum of squared residuals” (Introduction to Multiple Linear Regression, 2020).*

* **Justify the selection and development of the method (the “why”).**

*The method was chosen based on its ability to analyze and quantify relationships between multiple independent variables and a single dependent variable. This did exactly what was proposed and provided a prediction of a salary based on several variables. There is no other method that would fit better in this instance. This doesn’t mean that the method cannot be refined, and more variables could be added to ensure greater accuracy.*

*Another option that could have been used is a simple linear regression with just one independent variable and one dependent variable, but this would make for a useless model since the accuracy would be very low.*

## Validation

* **An appropriate validation method.**

*For validation, I used a test set with 20% of the cleaned data to evaluate the model’s performance. The main validation metrics used were Mean Squared Error and R-squared. MSE is a metric used to quantify the average squared difference between the estimated values and actual values, giving a picture of the model’s prediction accuracy. R-squared “is the proportion of the variance in the response variable that can be explained by the explanatory variables” (Introduction to Multiple Linear Regression, 2020) and provides a number on how good a fit the model is with the independent variables.*

* **Results of the validation method or a future plan to obtain those results.**

*The results indicated moderate accuracy levels suggesting that the model has potential to scale with more data. Although the accuracy wasn’t high, the results look promising for future applications with increased data volume This validation process serves as a foundation for further refining and enhancing the model’s predictive capabilities.*

## Visualizations

*The three visualizations are included in the notebook along with the tool:*

1. *A histogram of the salary ranges used in the algorithm.*
2. *A scatterplot of experience vs. salary.*
3. *A bar chart of average salaries of different company sizes.*

* Chart, histogram

  Description automatically generated
* Chart, line chart

  Description automatically generated
* Chart, bar chart

  Description automatically generated

## User Guide

1. Download the project zip folder with two notebooks and a csv file. Extract the files.
2. Navigate to [Project Jupyter | Home](https://jupyter.org/)
3. Click on “Try” on the top bar.

A picture containing diagram

Description automatically generated

1. Click on “JupyterLab”.

Graphical user interface, text, application, chat or text message

Description automatically generated

1. Click on the upload button.



1. Upload the file “C964\_Capstone\_Project\_JupyterLite”.
2. Repeat step 5 to additionally upload the file “ds\_salaries”.
3. Double-click on the “C964\_Capstone\_Project\_JupyterLite” file in JupyterLite.
4. Ensure that the kernel selected is Python.
5. Select Run -> Run All Cells
6. The visualizations and markdowns are provided above the tool. To use the tool, scroll to the bottom of the screen.
7. A work year can be selected with the slider.
8. The rest of the items are dropdown menus.
9. Select “Predict Salary”
10. The predicted salary will appear in the box below.

# Reference Page

Following APA guidelines, include references for any cited works, e.g., (Author, year). References are not required, and this page can be removed if no references are used. To cite sources used for code, you should include the references as code comments within the source code.

*Introduction to Multiple Linear Regression*. (2020, October 27). Retrieved from Statology: https://www.statology.org/multiple-linear-regression/