C964: Computer Science Capstone

Forecasting tool to predict the potential salary of recent graduates for the loan applicant profile of CannBank

Author: Daniel Santiago Sandoval Higuera

Student ID: 012366219

Submitted: February 27, 2025

Table of Contents:

Task 2 parts A, B, C and D

Part A: Letter of Transmittal

February 26, 2025

Ms. Michael Tana

Managing director of Loans

CannBank

2901 Street Can Maincan Road 3345. Office 402.

Subject: Proposal for a Salary Prediction Tool to Enhance Loan Application Process

Dear Ms. Tana,

As part of the CannBank Innovation Connect initiative, which looks for new ways to improve how we do business across CannBank, I am pleased to submit a proposal for a new forecasting tool that will predict the potential salary of recent college graduates, which can be used to improve CannBank's loan application process. The tool uses data-driven insights to estimate the income of job-seeking graduates or graduates who just got their first job, providing more accurate information for the creation of customer loan profiles.

Recent graduates represent a key group with the potential to become long-term, valuable customers as they are just starting to learn about the financial system. However, many of them are in the process of securing their first job, which can make obtaining financial support from banks challenging. Without a reliable income history, they often face difficulties in securing loans, which can hinder their ability to manage expenses or invest in their future. As part of CannBank's loan application system, determining an applicant's potential salary is a crucial factor in evaluating their loan eligibility. Currently, many recent graduates are entering the workforce without a clear salary benchmark, making it difficult to assess their loan risk. The absence of a reliable salary predictor for this group leads to a misunderstanding of their financial profile, potentially affecting both the bank's risk management and the graduates' ability to secure loans.

To help recent graduates enter the financial system in a more secure way, I am proposing to develop a forecasting tool based on 2021 National Survey of College Graduates data provided by the National Center for Science and Engineering Statistics (NCSES). By leveraging this data, the tool will predict the potential salary of recent graduates (looking for their first job or already in their first job) based on key variables such as degree type, field of study, age, marital status, employer type, etc. The tool will apply predictive models to generate salary estimates, which can then be integrated into CannBank's loan application system to assess graduates' loan profiles more accurately. As it is very important for the bank's equal opportunity policy, the tool will be free from bias, particularly in factors that had been historically unfair for some people, such as race and gender. Additionally, all data used will comply with privacy regulations, and no personally identifiable information will be used in the predictive model.

The salary predictions will provide a more data-driven and accurate measure of potential income for job-seeking graduates. Even If the data from a survey is from 2021, we can simulate the salary equivalency for 2025. With better salary forecasting, CannBank will be able to evaluate loan applications from recent graduates more effectively, reducing the risk of loan defaults.

Although timelines for the creation of this forecasting tool could be flexible depending on the time availability of your team, we estimate that it will take three months to deliver and start using it as a central part of the loan application process. This timeline includes loan application business understanding by data scientists from the innovation and research team, design of the solution, development of the predictive algorithm, evaluation, deployment, and testing of the final solution. An estimated cost for the development and initial adoption of the tool is around USD 100,000.

This is a unique opportunity for a collaboration between the Loan Application team and the Innovation and Research team led by me. We have extensive experience in data analysis, machine learning, and predictive modeling. Previously, we have worked on several successful projects involving teams all across the organization, such as the Logistics Team, with whom we were able to save USD$1’200.000 per year by implementing a better ATM-machine maintenance system, and with the service desk team, which led to an improvement of customer satisfaction of 20% and a reduction of service times by 32%. Personally, I am a computer scientist specializing in data science and implementing data-driven solutions. I started in the bank 10 years ago as a customer service associate and gradually started to move to more leadership positions; this experience allowed me to get a broad understanding of the bank.

I look forward to discussing how this project can move forward and benefit CannBank. Please feel free to reach out if you have any questions or require further information.

Sincerely,

Daniel Santiago Sandoval Higuera
Leader of the Innovation and Research Team
CannBank
Office 305
2901 Street Can Maincan Road 3345

Part B: Project Proposal Plan

Project Summary

This project will develop a forecasting tool for CannBank to predict the potential salary of recent graduates who do not yet have a job or just started their first job. This tool will utilize machine learning predictive models to analyze the 2021 National Survey of College Graduates dataset obtained from the National Center for Science and Engineering Statistics (NCSES). The predicted salary will serve as an input for CannBank's loan application system, allowing for better-informed financial decisions regarding loan approvals.

Currently, CannBank loan analysts do not have a structured method to analyze the potential income of a person just entering the labor force; this makes it difficult for some recent graduates to receive loans and bank products and start on the best terms of their financial relationship with the bank. To be able to assess risk and determine appropriate loan amounts, a reliable forecasting program will be developed to estimate future salaries for loan applicants who have recently graduated. The model's predictions will be one of CannBank's analysts' data used to assess risk and determine appropriate loan amounts.

This project will include the deliverable of a finished deployed prediction model that analysts will be able to access via a web application, a user guide, and a technical implementation report.

This project will be developed in conjunction with the Loan Application team and the Innovation and Research team. With this project, we expect to be able to evaluate loan applications from recent graduates more effectively, reducing the risk of loan defaults and ensuring a better experience with the bank for new clients that are just entering the loan financial system.

Data Summary

The data was obtained from the National Survey of College Graduates made in the year 2021 by the National Center for Science and Engineering Statistics (NCSES) website. https://ncses.nsf.gov/explore-data/microdata/national-survey-college-graduates

The survey includes data on the nation's college graduates, with a particular focus on those in the science and engineering workforce, nonetheless it include information from many different professions. It is sponsored by the National Science Foundation and conducted by the Census Bureau.

It comes as a ZIP file containing various files that are useful to understand the data, we will make use of the following:

- The `**epcg21.sas7bdat**` file contains the raw data in a compressed format, SAS format. (Column row values should be mapped with the information from Ppcg21.sas file).
- The `**Dpcg21.xlsx**` file contains an explanation of the meaning of the columns of the dataset (it has 537 different columns).

- The `**Ppcg21.sas**` file contains the meaning of the values inside every column row. (The values in **epcg21.sas7bdat**` are encoded to reduce the size of the dataset file at transmission)

These relevant files will be stored in the base_data folder within the designated drive space for this project. To request access, contact Karol Star at karol_s@cannbank.com, the data owner of these iniciatives. The data will be managed with proper documentation explaining its structure and usage, ensuring controlled access through an assigned owner. Describe how data will be processed and managed throughout the application development life cycle: design, development, maintenance, or others

Note that the information used is from a survey from 2021 because it is the last available survey on the National Center for Science and Engineering Statistics website. Maintainance of the system to use new survey data will be needed to ensure the system continues up to date.

Regular updates will be implemented to maintain data integrity and usability as new information becomes available. Version control practices will be followed to track changes, and periodic reviews will ensure the data remains accurate, relevant, and aligned with project objectives.

The data will be manipulated by using Python, specifically pandas. The raw data from `**epcg21.sas7bdat**` will be loaded, and its encoded row values will be decoded by using the data from the `**Ppcg21.sas**` file. To improve analysis readability, `**Dpcg21.xlsx**` will be read to get more descriptive column names and explanations.

Because the project will predict the yearly salaries of customers, the data will be first filtered to remove any rows that have missing salary information. It may be safe to assume that people just getting into the financial loan service will have a starting salary of less than USD $150.000. To ensure we do not include outliers or incorrectly input data by the survey respondents, we will include only information from respondents whose salary is less than USD $150.000. A review will be conducted to determine the number of values in each column that correspond to unanswered survey responses. If more than 30% of respondents did not answer a specific question, it will be considered for exclusion from the analysis. This will reduce the amount of columns available from the starting 537 columns.

The Fair Housing Act (FHA) and the Equal Credit Opportunity Act (ECOA) protect consumers by prohibiting unfair and discriminatory practices; this includes not discriminating by race, national origin, sex, or marital status (Office of the Comptroller of the Currency, n.d.). So, we will ensure not to include this information in the predictions made by the machine learning model. Note: Age is a significant factor for potential income, so it will be used to evaluate the potential income that may help determine the default risk of a loan. (it won't be used as an exclusion factor in the final decision by the analysts as they come to analyze the person's specific situation more in-depth before deciding about approving a loan).

Implementation

To implement the project, the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology will be used; CRISP-DM is a structured yet flexible methodology for data science projects; it has an iterative approach that ensures continuous improvement, making it a widely adopted framework for building reliable and practical data-driven solutions (Smart Vision Europe, n.d.). The following are the phases the project will include:

Business Understanding: Understand the CannBank requirements and the information they can ask their clients to make sure they align with the information used later in the modeling phase.

Data Understanding: Download, clean, and explore the 2021 National Survey of College Graduates Data files obtained from the National Center for Science and Engineering Statistics (NCSES) website. The descriptive method will be used here. Note: I have already done a quick exploration of the data to make sure I can generate a .csv file from the files NCSES provides and import the CSV in Python.

Data Preparation: Prepare the data for the machine learning models by handling missing values, encoding categorical data, and normalizing numerical features.

Modeling: Develop and train machine learning models to predict salaries. (Go back to data preparation as necessary)

Evaluation: Assess model performance using various evaluation metrics and select the best model. These metrics will be measured using cross-validation techniques such as k-fold cross-validation. The metrics could include Mean Absolute Error (MAE) Root Mean Square Error (RMSE) R-squared.

Note: Modeling and evaluation will be a single phase/deliverable to be apble to hyper-tune model parameters and select different models iteratively.

Deployment: Integrate the model into a web application, allowing bank analysts to input loan applicant details and obtain salary predictions.

At every phase of the process, proper documentation will be created.

Timeline

Milestone or deliverable	Duration (hours or days)	Projected start date	Anticipated end date
Business Understanding	1 week	1-Mar-25	7-Mar-25
Data Understanding	2 weeks	8-Mar-25	21-Mar-25
Data Preparation	3 weeks	22-Mar-25	13-Apr-25

Modeling and Evaluation	4 weeks	14-Apr-25	14-May-25
Deployment	2 weeks	15-May-25	30-May-25

Evaluation Plan

The project will be evaluated at the end of each Milestone as follows:

Milestone or deliverable	Evaluation method
Business Understanding	Completion of user stories and acceptance criteria to serve as reference points throughout development and approval from 100% of the stakeholders before proceeding to the next phase.
Data Understanding	Reports are generated with descriptive statistics and visualization to ensure data accuracy and representativeness and less than 30% missing or inconsistent data after processing. At least one graph of the interested features needs to be generated.
Data Preparation	Transformations to data required by the models are able to run without any problems.
Modeling and Evaluation	At least two different models are tested and compared. One model is chosen for deployment. Achieve a Mean Absolute Error of less than $22.000 in the predictions.
Deployment	Predictions take less than 3 seconds. Service is available as a web application and loan analyst are able to use it from their own computers without the need to install any sofware.

To consider successful the project all user stories and acceptance criteria need to be met.

Additionally, post-deployment and after monitoring for at least six months, a comparison within the model predictions with actual salary outcomes will be done to examine the effectiveness of the solution with real customers.

Resources and Costs

Hardware and Software Costs:

Data Scientist and software engineer rented computers and Google Collab virtual machines to run models. Estimated cost: $5,000.

Development Tools: Open-source tools will be used. Estimated cost: $0.

Labor Time and Costs:

1 Data Scientist: Responsible for data analysis, modeling, and evaluation. Estimated 400 hours at $100/hour, totaling $40,000.

1 Software Engineer: Handles web application development and deployment. Estimated 100 hours at $90/hour, totaling 9,000.

Project Manager: Ensures project is going in the right direction and coordinates. Estimated 40 hours at $110/hour, totaling $4,400.

Environment Costs:

Test deployment hosting: Hugging Face hosting resources for prototype development costs. Estimated at $1000 per month.

Maintenance: Regular updates and monitoring of the model and application. Estimated at $1,000 per month.

Hosting final web application: Cloud hosting services for the web application using AWS SageMaker services and EC2 instances. These services grow as service use increases. Estimated at $0.02/per inference.

Part C: Application

The code used for the implementation of the solution can be found in this Jupyter Notebook (Google Colab):

https://colab.research.google.com/drive/13w7AHwjKZVRQcnUY4n0dpz7nzUSW6dUG#scrollTo=169e14e4-6039-4261-b5ab-8f2febeaa657

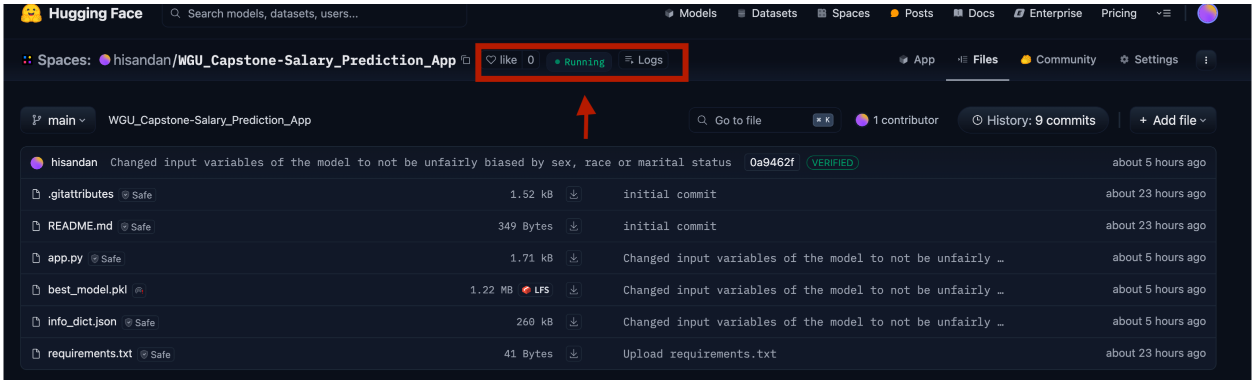
If the previous URL does not work, please use this:

https://drive.google.com/file/d/13w7AHwjKZVRQcnUY4n0dpz7nzUSW6dUG/view?usp=sharing

I encourage you to try it out as it contains all the processes of Crisp-DM to develop a prototype of the final product and what the development of the real potential salary prediction tool may look like. Note that a real-world application would require more analysis and tests.

Please use the 'User Guide' described in part D of this document to use the application as an online web application. (Go to: https://huggingface.co/spaces/hisandan/WGU_Capstone-Salary_Prediction_App)

To maintain the product, it will be hosted as a repository provided by Hugging Face (https://huggingface.co/spaces/hisandan/WGU_Capstone-Salary_Prediction_App/tree/main). Here, I will be able to analyze its Status and its Logs in a production environment. If changes are needed, they can be easily uploaded to this repository.



hisandan Changed input variables of the model to not be unfairly biased by sex, race or marital status 0a9462f VERIFIED about 5 hours ago

.gitattributes Safe	1.52 kB	initial commit	about 23 hours ago
README.md Safe	349 Bytes	initial commit	about 23 hours ago
app.py Safe	1.71 kB	Changed input variables of the model to not be unfairly …	about 5 hours ago
best_model.pkl	1.22 MB LFS	Changed input variables of the model to not be unfairly …	about 5 hours ago
info_dict.json Safe	260 kB	Changed input variables of the model to not be unfairly …	about 5 hours ago
requirements.txt Safe	41 Bytes	Upload requirements.txt	about 23 hours ago

Finally, the functional dashboard that includes 3+ visualization types implementation can be found inside the Jupyter Notebook with the implementation of the solution (https://colab.research.google.com/drive/13w7AHwjKZVRQcnUY4n0dpz7nzUSW6dUG#scrollTo=169e14e4-6039-4261-b5ab-8f2febeaa657) these graphs are for internal purpose only (developers of the product and analyst interested on trends on the income) so they are not included as an additional specific dashboard or as a web application.

Part D: Post-implementation Report

Solution Summary

CannBank lacked a structured method to estimate the potential income of recent graduates applying for loans. Without an accurate way to assess earning potential, loan approvals become difficult, potentially limiting financial opportunities for new graduates and increasing risk for the bank.

To address this issue, a machine learning-powered salary prediction tool was developed. This tool analyzes data from the 2021 National Survey of College Graduates (NSCG) and predicts potential salaries based on factors such as field of education, expected company type and size where the graduate will work, and other relevant attributes. The predictions will serve as input for CannBank's loan application system, helping analysts make informed decisions on loan approvals and risk assessment.

The prediction tool has been made available to loan analysts as a web application that can be found here: https://huggingface.co/spaces/hisandan/WGU_Capstone-Salary_Prediction_App

Data Summary

The raw data came from the National Survey of College Graduates (NSCG) 2021, provided by the National Center for Science and Engineering Statistics (NCSES). It is publicly available at: https://ncses.nsf.gov/explore-data/microdata/national-survey-college-graduates

The file downloaded is https://ncses.nsf.gov/822/assets/0/files/college_grads_2021.zip. This zip was extracted and contains multiple files with the data and with information to understand the data and its recollection.

The data file comes in a SAS format, which can be loaded directly into Python by using pandas.

To understand the data, we need to take into account the following files included inside the extracted ZIP folder from the download step:

- The `epcg21.sas7bdat` contains the data.

- The `Dpcg21.xlsx` file contains an explanation of the meaning of the columns of the data frame.

- The `Ppcg21.sas` file contains the meaning of the values inside every column row. (The values loaded from `epcg21.sas7bdat` were encoded to reduce the size of the transmission file)

Because the data obtained had 537 different columns relevant to the business case, a subset of columns was selected as potential variables to use. A manual selection of the columns of interest to use based on the column explanations from the Dpcg21.xlsx file and my bank business understanding was made. If a column had more than 30% missing values, the column was discarded. To eliminate outliers only rows with annual salary under $150,000 were used. Finally, features related to race, gender, and marital status were excluded in compliance with Fair Lending regulations.

To prepare the data for machine learning models, the numerical data was imputed with the mean and standardized, and the categorical data was imputed with the most frequent value and one-hot encoded.

Finally, the dataset was split into training and testing sets using an 60-20-20 split to enable model training, evaluation and validation.

Machine Learning

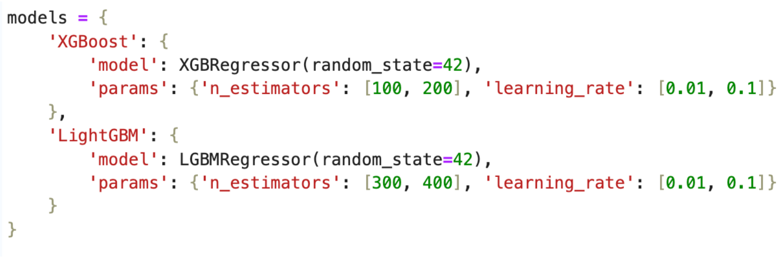
Gradient boosting is a machine learning algorithm used for classification and regression tasks. It builds a strong model by combining many weak models, usually decision trees. It works by training these trees one after another, where each new tree focuses on correcting the mistakes made by the previous ones. This gradual improvement process is why it’s called "boosting". (Geeks for Geeks, 2023)

Different Gradient boosting models were used: XGBoost and LightGBM; these models are able to use numeric and categorical inputs and produce a numeric prediction. Hyperparameter tuning was performed using grid search for n_estimators and learning_rate.

XGBoost (Extreme Gradient Boosting) is an ensemble learning method that improves predictions by combining multiple decision trees with gradient boosting. It is highly efficient for structured datasets.

LightGBM (Light Gradient Boosting Machine) is a gradient-boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient.

The following image shows the models and the combination of hyperparameters used by each model:

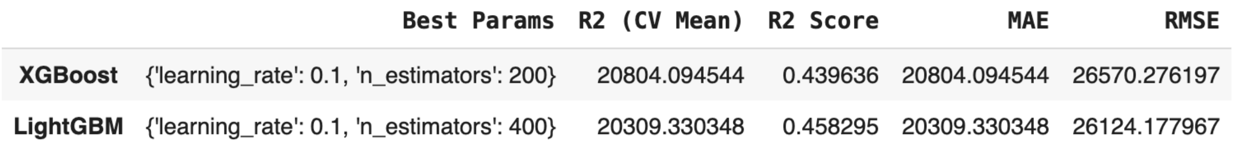


```
models = {
    'XGBoost': {
        'model': XGBRegressor(random_state=42),
        'params': {'n_estimators': [100, 200], 'learning_rate': [0.01, 0.1]}
    },
    'LightGBM': {
        'model': LGBMRegressor(random_state=42),
        'params': {'n_estimators': [300, 400], 'learning_rate': [0.01, 0.1]}
    }
}
```

These models were selected because they are applicable to structured datasets with categorical features and regression tasks. LightGBM, the model selected to go to the production environment, can be distributed, has Lower memory usage, and is capable of handling large-scale data.

Note that an even stronger implementation of the prediction algorithm can try more different types of models and select the best-performing one.

In the following table, we can observe that the best-performing model is the LightGBM, with a average error in predictions (MAE) of $20.309



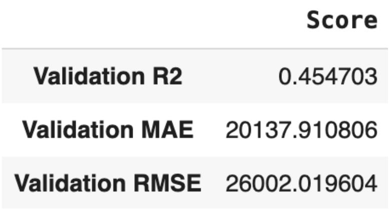
	Best Params	R2 (CV Mean)	R2 Score	MAE	RMSE
XGBoost	{'learning_rate': 0.1, 'n_estimators': 200}	20804.094544	0.439636	20804.094544	26570.276197
LightGBM	{'learning_rate': 0.1, 'n_estimators': 400}	20309.330348	0.458295	20309.330348	26124.177967

Validation

The goal is to find the best-performing model based on **mean absolute error (MAE).** The mean absolute error (MAE) gives us a sense of the average error of the salary predictions; for example, if the predicted salary is `$100.000` is probable that the real potential salary that the loan applicant will get is between `$80.000` and `$120.000`.

The hyperparameters were tested iteratively to find the best-performing model of each type, measuring performance as the least mean absolute error. After this we found that LightGBM (Light Gradient Boosting Machine) algorithm with a learning rate of 0.1 and 400 estimators is the best performing model.

A new group of data (validation data) was used to generate the final metrics to validate that the algorithm is not overfitting the data. The following are the validation metrics of the best algorithm in the validation data:



	Score
Validation R2	0.454703
Validation MAE	20137.910806
Validation RMSE	26002.019604

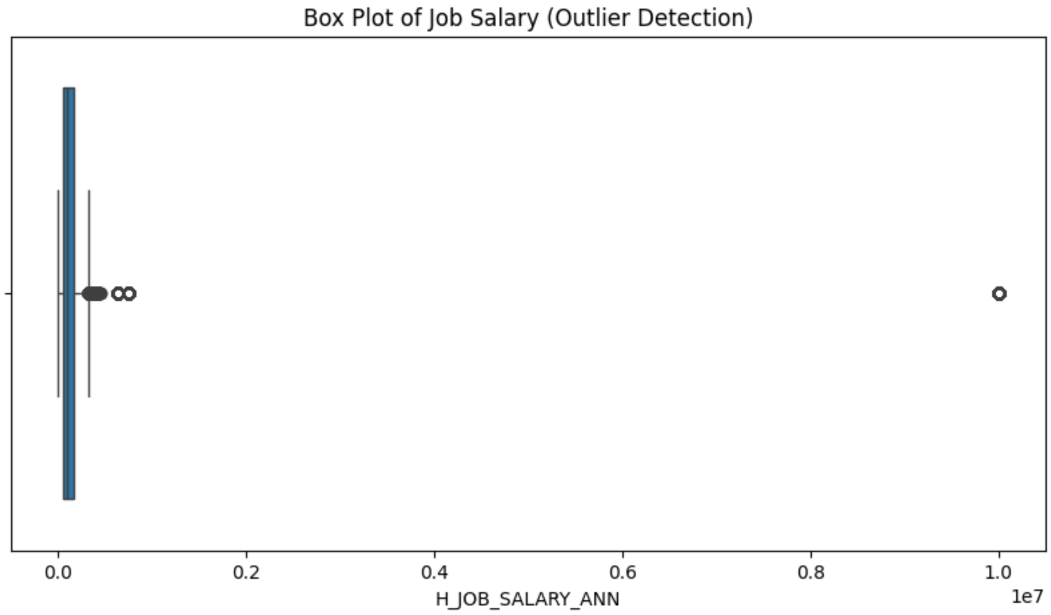
Note that for the purpose of the analysis, the only metric used was Mean Absolute Error (MAE) as it is very straightforward to understand. In this case, the average error of the prediction of yearly salary is USD $20.137 per year.

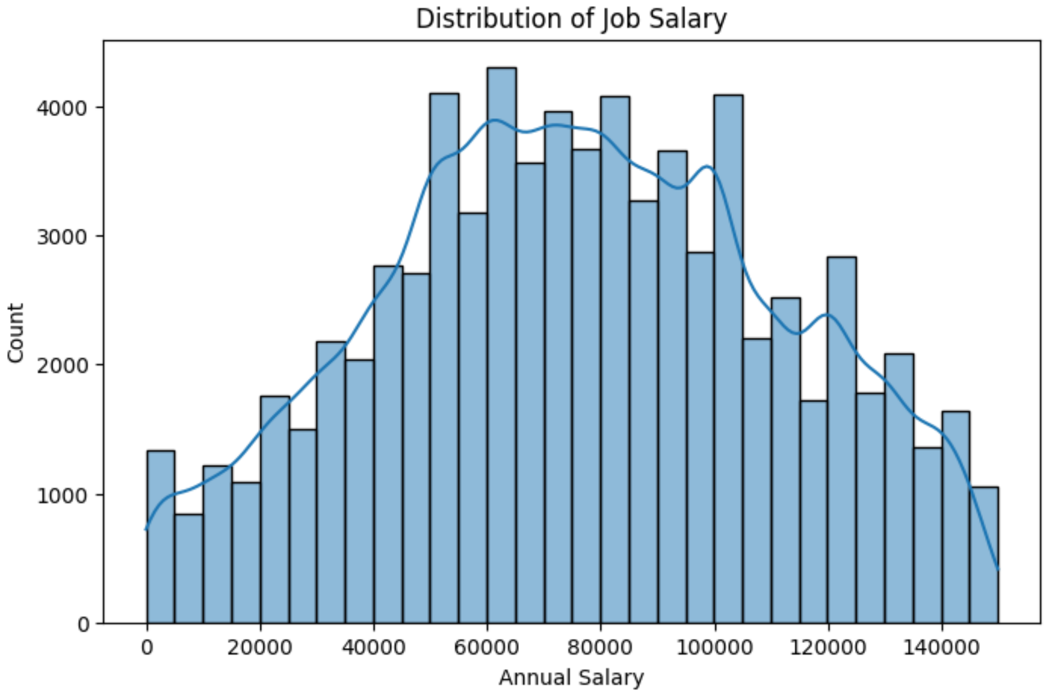
A comparison within the model predictions with actual salary outcomes can be done in the future to examine the effectiveness of the solution with real customers.

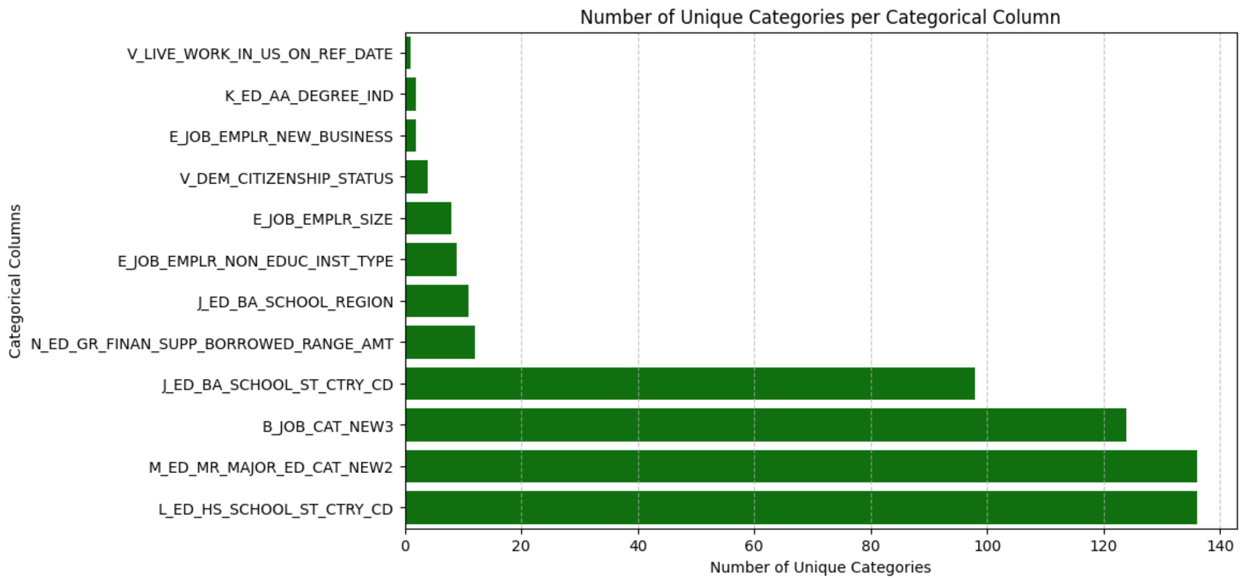
Visualizations

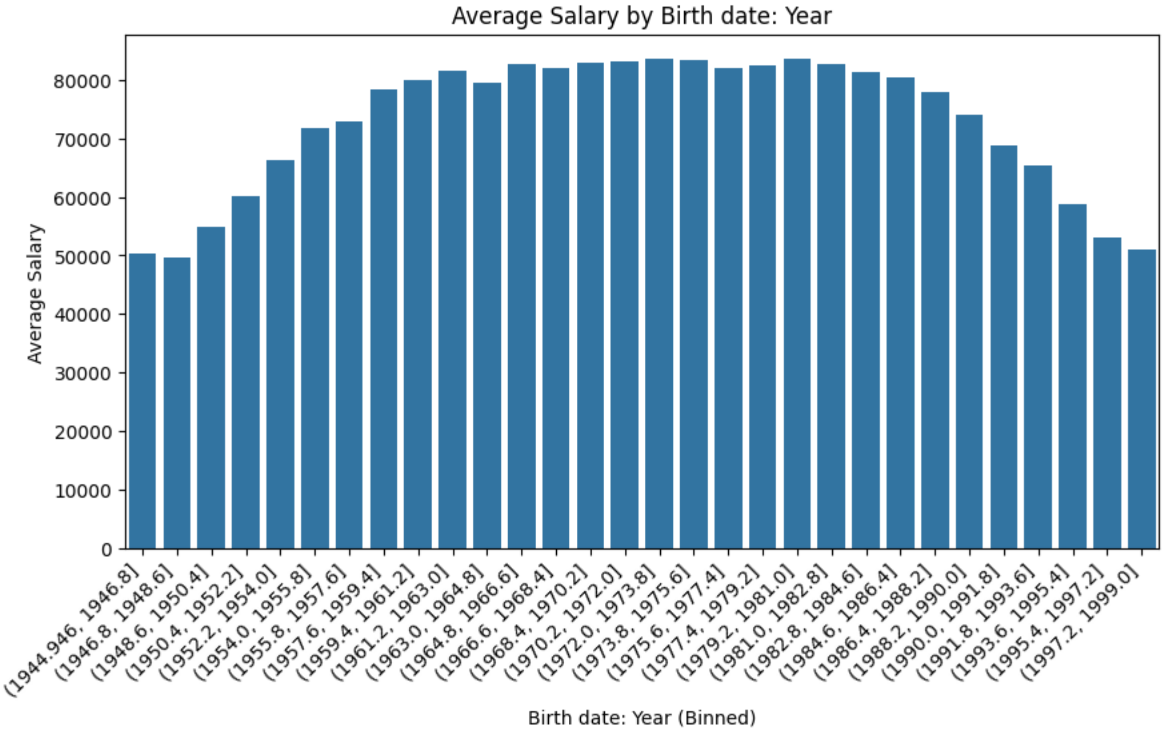
To access the visualizations, open the following link and scroll down through the "Phase 2: Data Understanding" section: https://colab.research.google.com/drive/13w7AHwjKZVRQcnUY4n0dpz7nzUSW6dUG#scrollTo=05d9504b-113f-4ebf-bc15-3374ce52e8c4

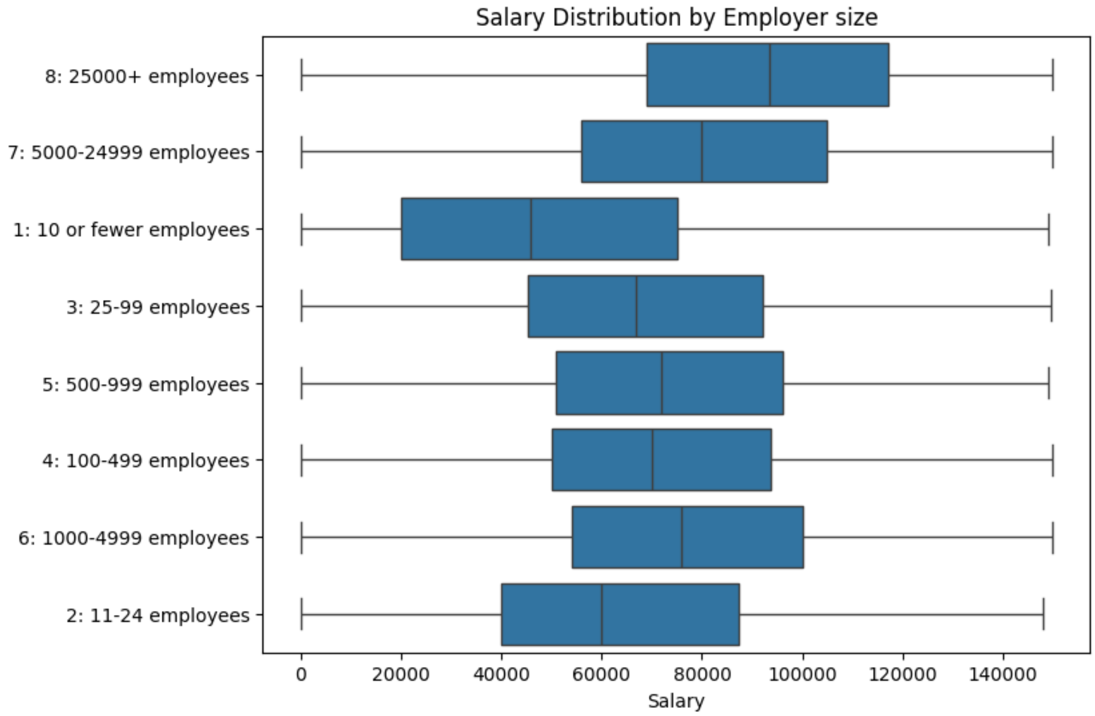
The following are some of the most interesting visualizations obtained while exploring the Census data:

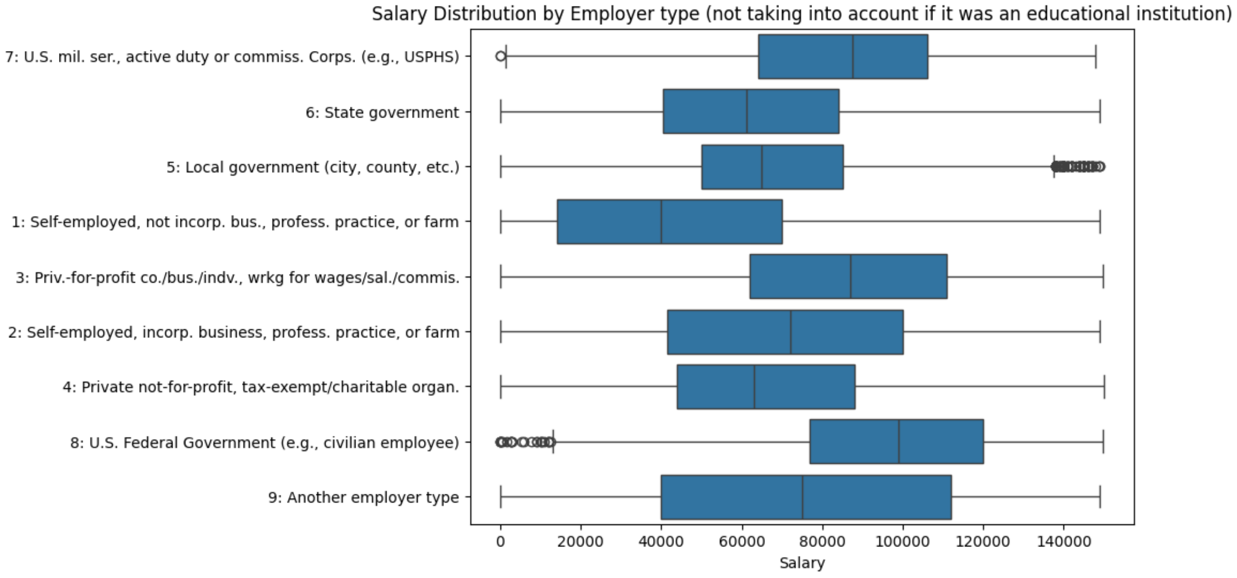

Box Plot of Job Salary (Outlier Detection)
0.0
0.2
0.4
0.6
0.8
1.0
1e7
H_JOB_SALARY_ANN


Distribution of Job Salary
Count
0
1000
2000
3000
4000
0
20000
40000
60000
80000
100000
120000
140000
Annual Salary


Number of Unique Categories per Categorical Column
Categorical Columns
V_LIVE_WORK_IN_US_ON_REF_DATE
K_ED_AA_DEGREE_IND
E_JOB_EMPLR_NEW_BUSINESS
V_DEM_CITIZENSHIP_STATUS
E_JOB_EMPLR_SIZE
E_JOB_EMPLR_NON_EDUC_INST_TYPE
J_ED_BA_SCHOOL_REGION
N_ED_GR_FINAN_SUPP_BORROWED_RANGE_AMT
J_ED_BA_SCHOOL_ST_CTRY_CD
B_JOB_CAT_NEW3
M_ED_MR_MAJOR_ED_CAT_NEW2
L_ED_HS_SCHOOL_ST_CTRY_CD
0
20
40
60
80
100
120
140
Number of Unique Categories


Average Salary by Birth date: Year
Average Salary
80000
70000
60000
50000
40000
30000
20000
10000
0
(1944.946, 1946.8]
(1946.8, 1948.6]
(1948.6, 1950.4]
(1950.4, 1952.2]
(1952.2, 1954.0]
(1954.0, 1955.8]
(1955.8, 1957.6]
(1957.6, 1959.4]
(1959.4, 1961.2]
(1961.2, 1963.0]
(1963.0, 1964.8]
(1964.8, 1966.6]
(1966.6, 1968.4]
(1968.4, 1970.2]
(1970.2, 1972.0]
(1972.0, 1973.8]
(1973.8, 1975.6]
(1975.6, 1977.4]
(1977.4, 1979.2]
(1979.2, 1981.0]
(1981.0, 1982.8]
(1982.8, 1984.6]
(1984.6, 1986.4]
(1986.4, 1988.2]
(1988.2, 1990.0]
(1990.0, 1991.8]
(1991.8, 1993.6]
(1993.6, 1995.4]
(1995.4, 1997.2]
(1997.2, 1999.0]
Birth date: Year (Binned)





Guides

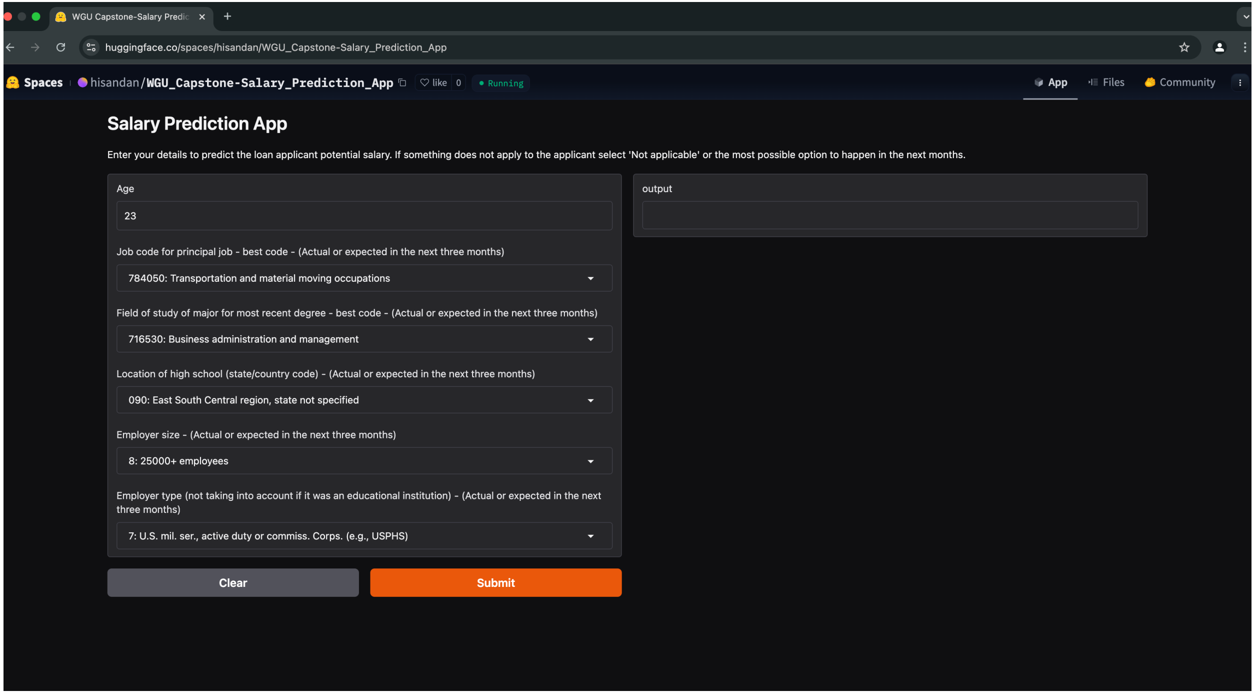
Visualizations Guide: (For internal loan analysts and data scientists)

To access the visualizations, open the following link and scroll down through the "Phase 2: Data Understanding" section: https://colab.research.google.com/drive/13w7AHwjKZVRQcnUY4n0dpz7nzUSW6dUG#scrollTo=05d9504b-113f-4ebf-bc15-3374ce52e8c4

Final product User Guide:

The use of the prediction tool of the potential salary of recent graduates for the loan applicant profile of CannBank is very straightforward, please use the following instructions:

1. Open this link in your preferred navigator: https://huggingface.co/spaces/hisandan/WGU_Capstone-Salary_Prediction_App


WGU Capstone-Salary Predic
huggingface.co/spaces/hisandan/WGU_Capstone-Salary_Prediction_App
Spaces
hisandan/WGU_Capstone-Salary_Prediction_App
like 0
Running
App
Files
Community
Clear
Submit

Salary Prediction App

Enter your details to predict the loan applicant potential salary. If something does not apply to the applicant select 'Not applicable' or the most possible option to happen in the next months.

Age: 23

Job code for principal job - best code - (Actual or expected in the next three months): 784050: Transportation and material moving occupations

Field of study of major for most recent degree - best code - (Actual or expected in the next three months): 716530: Business administration and management

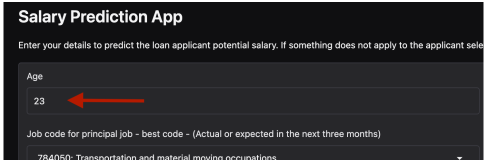
Location of high school (state/country code) - (Actual or expected in the next three months): 090: East South Central region, state not specified

Employer size - (Actual or expected in the next three months): 8: 25000+ employees

Employer type (not taking into account if it was an educational institution) - (Actual or expected in the next three months): 7: U.S. mil. ser., active duty or commiss. Corps. (e.g., USPHS)

output

2. Enter the age of the loan applicant in the Age box.



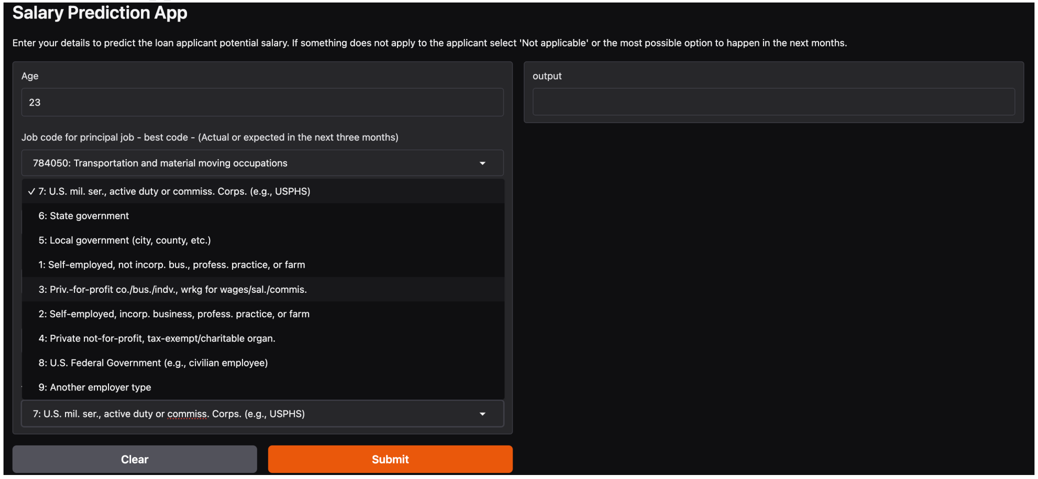
Salary Prediction App

Enter your details to predict the loan applicant potential salary. If something does not apply to the applicant sele

Age: 23

Job code for principal job - best code - (Actual or expected in the next three months): 784050: Transportation and material moving occupations

3. Use the dropdowns to make the most appropriate selections from an Actual or Expected situation of the loan applicant.


✓ 7: U.S. mil. ser., active duty or commiss. Corps. (e.g., USPHS)
6: State government
5: Local government (city, county, etc.)
1: Self-employed, not incorp. bus., profess. practice, or farm
3: Priv.-for-profit co./bus./indv., wrkg for wages/sal./commis.
2: Self-employed, incorp. business, profess. practice, or farm
4: Private not-for-profit, tax-exempt/charitable organ.
8: U.S. Federal Government (e.g., civilian employee)
9: Another employer type
Clear
Submit

Salary Prediction App

Enter your details to predict the loan applicant potential salary. If something does not apply to the applicant select 'Not applicable' or the most possible option to happen in the next months.

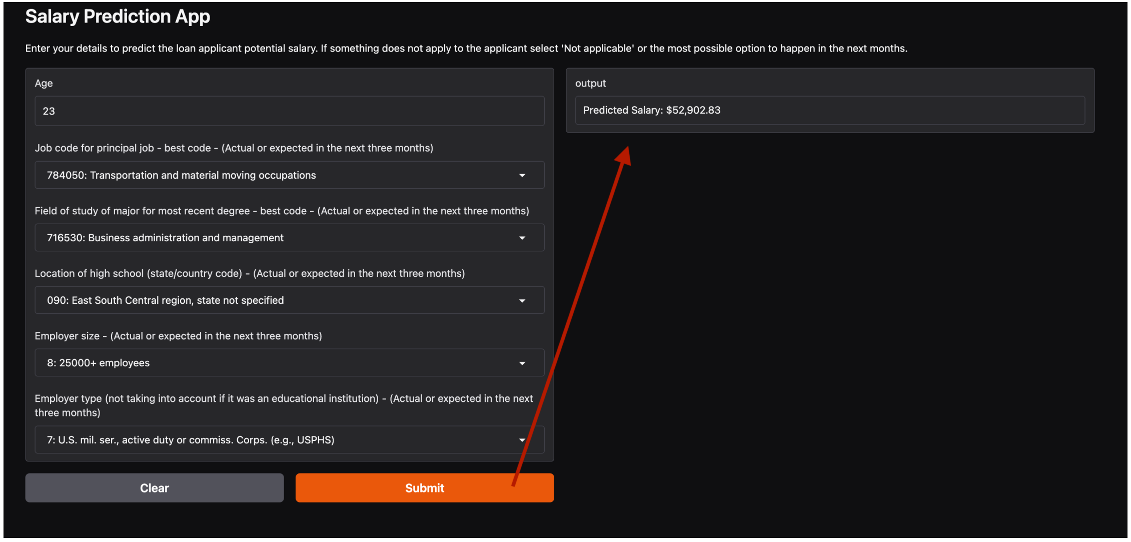
Age: 23

Job code for principal job - best code - (Actual or expected in the next three months): 784050: Transportation and material moving occupations

7: U.S. mil. ser., active duty or commiss. Corps. (e.g., USPHS)

output

4. Click submit to generate the potential salary prediction:



Salary Prediction App

Enter your details to predict the loan applicant potential salary. If something does not apply to the applicant select 'Not applicable' or the most possible option to happen in the next months.

Age: 23

Job code for principal job - best code - (Actual or expected in the next three months): 784050: Transportation and material moving occupations

Field of study of major for most recent degree - best code - (Actual or expected in the next three months): 716530: Business administration and management

Location of high school (state/country code) - (Actual or expected in the next three months): 090: East South Central region, state not specified

Employer size - (Actual or expected in the next three months): 8: 25000+ employees

Employer type (not taking into account if it was an educational institution) - (Actual or expected in the next three months): 7: U.S. mil. ser., active duty or commiss. Corps. (e.g., USPHS)

output: Predicted Salary: $52,902.83

References

- Office of the Comptroller of the Currency. (n.d.). Fair lending. U.S. Department of the Treasury. https://www.occ.treas.gov/topics/consumers-and-communities/consumer-protection/fair-lending/index-fair-lending.html
- Smart Vision Europe. (n.d.) What is the CRISP-DM methodology? https://www.sv-europe.com/crisp-dm-methodology/
- Geeks for Geeks. (2023) Gradient Boosting in ML. https://www.geeksforgeeks.org/ml-gradient-boosting/