DATA 606: CAPSTONE PROJECT IDEAS- Group A

[Simran Shah](mailto:sshah8@umbc.edu)

[Gelareh Vakili](mailto:av10299@umbc.edu)

[Dereck Román Rosario](mailto:dromanr1@umbc.edu)

**Date:** 06/10/2025

### **Salary Prediction Based on Education, Field, Location, and Experience**

### **Project Overview**

### In this capstone project, we aim to build a regression model that predicts a developer’s annual salary using individual-level data on education, job function, country of employment, and years of professional experience. Our goal is to analyze and quantify how these features affect compensation, and to identify which combinations of experience and educational attainment lead to higher-paying jobs across global regions.

### This project aligns with real-world workforce analytics applications and reflects challenges in career equity, compensation transparency, and hiring fairness, particularly in the evolving global tech industry.

### **Primary Dataset: Stack Overflow Developer Survey 2024**

* **Source:** Stack Overflow
* **Link:**<https://insights.stackoverflow.com/survey>
* **Direct Download:** <https://survey.stackoverflow.co/datasets/stack-overflow-developer-survey-2024.zip>
* **Format:** CSV (within ZIP file)
* **Size:** ~85MB uncompressed
* **Responses:** ~89,000 developers

**Attributes Used:**

* ConvertedCompYearly – (Target) total annual compensation
* EdLevel – highest level of education completed
* DevType – job role/field (e.g., backend dev, data scientist)
* Country – geographic location of employment
* YearsCodePro – years of professional coding experience
* Optional: RemoteRatio, Employment, LanguageHaveWorkedWith for secondary insights

**Planned Analysis:**We aim to predict developers’ annual compensation using regression analysis based on factors such as experience, education, and job role. We'll also explore correlations between various developer attributes and their income levels globally.

**Proposed Methods:**

* **Preprocessing:** Cleaning, handling missing values, encoding categorical variables
* **Modeling:** Linear Regression, Random Forest Regressor, XGBoost
* **Visualization:** Salary distributions, feature importance, country-wise trends

**Evaluation Metrics:**

* Root Mean Squared Error (RMSE)
* Mean Absolute Error (MAE)
* R² Score

**Why This Dataset?**It is detailed, well-structured, relevant to real-world hiring trends, and publicly available outside of Kaggle, making it compliant with course guidelines. It also provides individual-level records for robust modeling.

### **Backup Dataset: U.S. Bureau of Labor Statistics (OEWS 2023)**

* **Source:** U.S. Bureau of Labor Statistics (BLS)
* **Access Link:** <https://www.bls.gov/oes/tables.htm>
* **Direct Download**: <https://www.bls.gov/oes/special-requests/oesm24all.zip>
* **Format:** XLSX and CSV
* **Size:** Lightweight tables per occupation and state

**Key Attributes:**

* Job Title
* Mean and Median Salary
* Location (state and region)
* Industry codes

**Planned Use:**In case the primary dataset poses challenges, we will use BLS data to perform statistical analysis or clustering. We’ll identify wage trends across regions and industries, and group job types by compensation.

**Alternative Analyses:**

* Clustering job roles using salary and location
* Regional wage distribution visualizations
* Inflation-adjusted wage comparison

**Backup Metrics:**

* R² Score for regression
* Silhouette Score or Davies–Bouldin Index for clustering