COLLEGE OF APPLIED BUSINESS AND TECHNOLOGY

(Affiliated to Tribhuvan University)

Report on Salary Prediction System

Submitted by:

Arnav Sharma (103)

Kushal Basnet (113)

Lotus Kshetri (114)

Submitted to:

CAB-Tech

COLLEGE OF APPLIED BUSINESS AND TECHNOLOGY

(Chabahil, Gangahity, Kathmandu)

CERTIFICATE OF AUTHENTICATION

This is to verify that *Mr. Arnav Sharma, Mr. Kushal Basnet and Mr. Lotus Kshetri*, students of B.Sc.-CSIT fourth semester, have submitted assignment report for their practical fulfillment of EPC subject **Machine Learning** project.

.....

Subject Teacher
Tekendra Nath Yogi

Table of Contents

1. Introduction

1.1 Background and Motivation

The increasing global demand for skilled software developers has led to significant variations in their salaries across different countries and based on their work experience. This project aims to develop a machine learning model that predicts software developers' salaries worldwide and evaluates the results considering both the country and the work experience of the developers.

The demand for software developers has increased, leading to diverse job opportunities worldwide. Companies seeking to hire software developers from across the globe can benefit from this project by understanding the salary expectations of developers in different regions. The conventional approach of analyzing salary data based on average or median figures may not provide accurate insights due to the underlying heterogeneity of data points. This information can assist governments and organizations in shaping policies and initiatives to foster technological growth and innovation.

1.2 Objectives

The primary objectives of the multiple disease prediction app are as follows:

- Create a regression-based ML model for global software developer salary prediction, integrating experience and education.
- Assess model accuracy using appropriate metrics, aiding companies in aligning compensation with market norms.
- Regularly update the model with new data to enhance accuracy and relevance over time.

2. Dataset and Preprocessing

2.1Description of Datasets Used for the Predictions

A structured collection of data containing information about software developers from around the world. The dataset's purpose is to train a predictive model that can accurately estimate software developers' salaries based on their country and work experience.

- **Salary Prediction Dataset:** The salary prediction models dataset includes attributes such as country name, education level, experience in years, employment status, and corresponding salary.

Country	EdLevel	YearsCodePro	Employment	Salary
Canada			Employed, full-time	
United Kingdom of Great Britain and Northern Ireland	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	5	Employed, full-time	40205
Israel	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	17	Employed, full-time	215232
United States of America	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	3	Employed, full-time	
Germany	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)		Student, full-time	
India	ool (e.g. American high school, German Realschule or Gymnasium, etc.)		Student, part-time	
India	Some college/university study without earning a degree		Not employed, but looking for work	
Netherlands	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	6	Employed, full-time	49056
Croatia	Some college/university study without earning a degree	30	Independent contractor, freelancer, or self-employed	
United Kingdom of Great Britain and Northern Ireland	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	2	Employed, full-time	60307
United States of America	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	10	Employed, full-time;Independent contractor, freelancer, or self-employed	194400
United States of America	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	5	Employed, full-time	65000
Australia	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	15	Employed, part-time	
United States of America	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	5	Employed, full-time;Independent contractor, freelancer, or self-employed	110000
Russian Federation	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	4	Independent contractor, freelancer, or self-employed	
Czech Republic	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	4	Employed, full-time	19224
Austria	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	10	Employed, full-time	202623
Austria	ool (e.g. American high school, German Realschule or Gy	22	Employed, full-time	51192

Country: A categorical variable representing the name of the country where the software developer is employed.

EdLevel: A categorical variable denoting the highest educational qualification of the software developer.

YearsCodePro: A numerical variable representing the number of years of work experience the software developer has in the field of software development.

Employment: A categorical variable indicating the current employment status of the software developer.

Salary: A numerical variable representing the salary of the software developer.

2.2 Data Preprocessing and Feature Engineering

Data preprocessing and feature engineering are essential steps in any machine learning project, including predicting salaries for software developers based on work experience and education level. These steps help improve the quality of data and create meaningful features that can better represent the underlying patterns in the data.

- **Data Cleaning:** We checked for missing data in the dataset and decided how to handle it. Missing values are imputed, removed, or replaced with suitable measures like mean, median, or mode based on the data distribution and context.
- **Feature Selecting:** We analyze the importance of features and select the most relevant ones for training the model. In this case, years of experience and education level are likely to be important features for salary prediction.
- **Feature Engineering:** Domain knowledge of the factors required for salary guided the selection of relevant features. We engineered new features, such as annual salary in dollars, that combined multiple attributes to improve the predictive power of the models.

3. Model Training

3.1 Decision Tree Regression algorithm

The first step is to load and preprocess the dataset containing information such as country names, years of experience, and education levels. Data preprocessing involves handling missing values, encoding categorical variables, and converting years of experience to numerical format if necessary.

Next, the dataset is split into a training set and a testing set. The training set is used to train the Decision Tree Regression model, while the testing set is used to evaluate its performance.

Once the model is instantiated, it is trained on the training data, where the features are the years of experience, education level, and country name, and the target variable is the salary.
After the model is trained, it can be used to make predictions on the testing set to estimate the salaries of software developers. Evaluation metrics like mean squared error, mean absolute error, or R-squared can be used to assess the model's accuracy in predicting salaries.
Applying the Decision Tree Regression algorithm to this dataset allows us to capture nonlinear relationships between the input features and the target variable, making it suitable for predicting salaries based on a combination of years of experience, education level, and country of work.

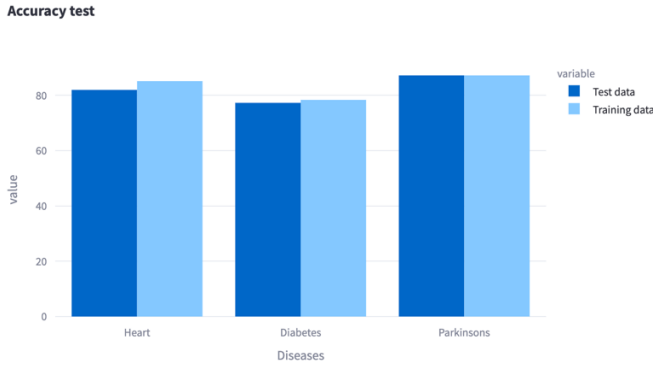
3.2 Linear Regression for Salary Prediction System

In this machine learning project using scikit-Learn's Linear Regression, the goal is to predict the salaries of software developers worldwide based on their work experience and education level. The dataset contains information on country names, years of experience, and education levels. After loading and preprocessing the data, including data cleaning, feature engineering, and encoding categorical variables, the dataset is split into training and testing sets. The Linear Regression model is trained using the training data and then used to make predictions on the testing data. Performance metrics like mean squared error, mean absolute error, and R-squared are calculated to evaluate the model's accuracy in predicting salaries.

4. Evaluation and Performance

4.1 Evaluation Metrics for Salary Prediction Model

Evaluating the performance of our disease prediction models is essential to determine their accuracy and effectiveness. So, we set the bar graph to determine whether our model is affected or not and, we determine the accuracy.



4.2 Insights from Model Evaluation and Analysis

Upon analyzing the model's predictions, we observed that certain features, such as age, gender, and specific medical attributes, played significant roles in disease prediction. Additionally, we investigated the effect of feature importance on model performance. The insights gained from this analysis can help guide future model improvements and feature selection.

5. Validation

Since the prediction of salary is not an exact science, for validation purpose, we do not have exact method for validation. But for the sake of validation, we compared the generated results with similar results found on the web and found little difference.

6. Streamlit Interface Integration

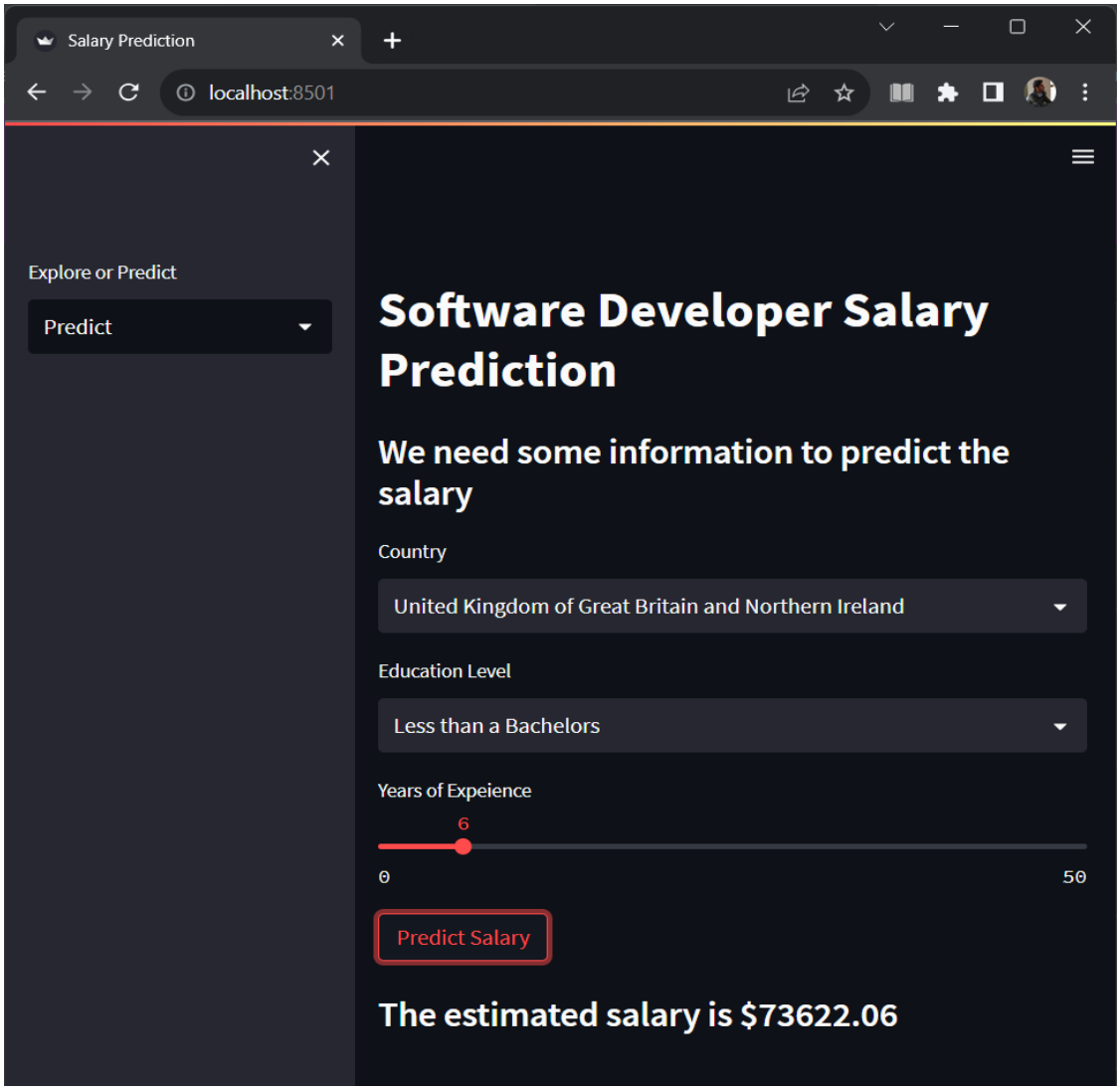
Streamlit is an open-source Python library that provides an intuitive and simple way to build interfaces for machine learning models. It supports a wide range of model types, including image classification, text generation, object detection, and more. Streamlit's key features include:

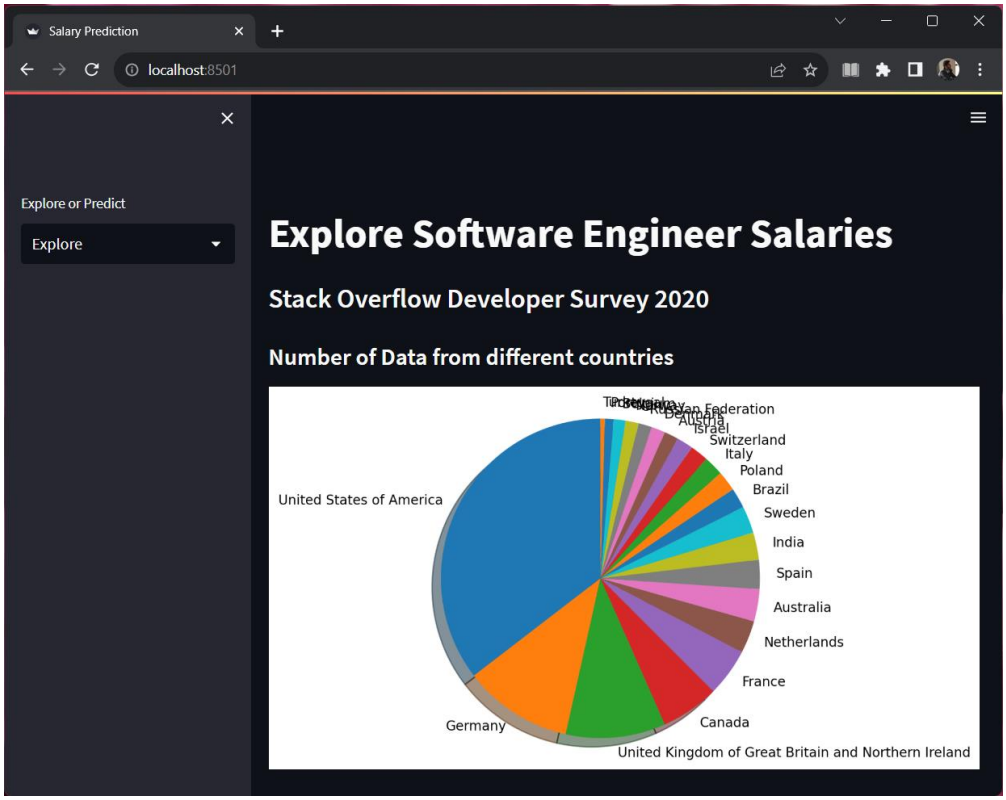
- **Rapid Prototyping**: Streamlit enables fast and easy development of interactive web applications directly from Python scripts. Developers can focus on their data and logic without worrying about front-end code.
- **Wide Range of Widgets**: Streamlit provides a variety of widgets (text inputs, sliders, buttons, etc.) that allow users to interact with the application and modify data dynamically.
- **Data Visualization**: Streamlit integrates seamlessly with popular data visualization libraries like Matplotlib, Plotly, and Altair, making it easy to display interactive plots and charts.
- **Sharing and Deployment**: Streamlit applications can be easily shared with others by simply sharing the Python script. Deployment to various platforms, including Heroku, AWS, or Streamlit Sharing, is also straightforward.

7. Hosting Web-App Locally

7.1 Deploying the Model as an API

The trained Salary Prediction for developers' model is hosted on Streamlit's API platform, making it easily accessible as an API service. Using streamlit to run the app, we can easily host the web app in the local host of the computer after installing all the requirements.




Salary Prediction
localhost:8501
Explore or Predict
Explore
Explore Software Engineer Salaries
Stack Overflow Developer Survey 2020
Number of Data from different countries
United States of America
Switzerland
Italy
Poland
Brazil
Sweden
India
Spain
Australia
Netherlands
France
Canada
Germany
United Kingdom of Great Britain and Northern Ireland


Salary Prediction
localhost:8501
Explore or Predict
Explore
Mean Salary Based On Country
150,000
100,000
50,000
0
Australia
Austria
Belgium
Brazil
Canada
Denmark
France
Germany
India
Israel
Italy
Netherlands
Norway
Poland
Portugal
Russian Federation
Spain
Sweden
Switzerland
Turkey
United Kingdom ...
United States of A...
Salary
Mean Salary Based On Experience
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
0
5
10
15
20
25
30
35
40
45
50

8. Conclusion

The software developer salary prediction app stands as a significant accomplishment, harnessing facial biometrics and deep learning to forecast salaries within the software development industry. Through its precise and user-friendly salary prediction results, the app possesses the potential to influence both job seekers and employers, giving valuable insights into salary projections. Its impact transcends individual users, reaching into the realm of software development job seeking and the industry at large. With almost accurate salary predictions and illuminating compensation prospects, the app can facilitate hiring protocols, enhance negotiating capabilities for job seekers, and contribute to a more transparent software development job market. In conclusion, the development of “the software developer prediction” app has been a rewarding journey, combining advancements in AI with the mission of helping people like us have a realistic expectation of their future. We look forward to further refining and expanding the app.

9. References

1. Stack Overflow Developer Survey 2020. (n.d.). Stack Overflow. https://insights.stackoverflow.com/survey/2020
2. Streamlit • A faster way to build and share data apps. (n.d.). https://streamlit.io/
3. Srmaarnav. (n.d.). *GitHub - srmaarnav/Salary_Prediction*. GitHub.

   https://github.com/srmaarnav/Salary_Prediction

Regression Analysis

4. Wikipedia contributors. (2023). Regression analysis. *Wikipedia*.

   https://en.wikipedia.org/wiki/Regression_analysis

5. GeeksforGeeks. (2023). Linear Regression Python Implementation. *GeeksforGeeks*.

   https://www.geeksforgeeks.org/linear-regression-python-implementation/

10. Appendix

The main code for app.py:

```
import streamlit as st
st.set_page_config(
        page_title="Salary Prediction",
)
from predict_page import show_predict_page
from explore_page import show_explore_page
page = st.sidebar.selectbox("Explore or Predict", ("Predict", "Explore"))
if page == "Predict":
    show_predict_page()
elif page == "Explore":
    show_explore_page()
```

The code for predict_page.py:

```
import streamlit as st
import numpy as np
import pickle
def load_model():
    with open('Pickle\saved_steps.pkl', 'rb') as file:
        data = pickle.load(file)
    return data
data = load_model()
regressor = data["model"]
le_country = data["le_country"]
le_education = data["le_education"]
def show_predict_page():
    st.title("Software Developer Salary Prediction")
    st.write("### We need some information to predict the salary")
      countries = (
        'United Kingdom of Great Britain and Northern Ireland',
```

```
        'Israel',
        'Netherlands',
        'United States of America',
        'Austria',
        'Italy',
        'Canada',
        'Germany',
        'Poland',
        'Norway',
        'France',
        'Sweden',
        'Spain',
        'Belgium',
        'India',
        'Brazil',
        'Switzerland',
        'Denmark',
        'Australia',
        'Portugal',
        'Russian Federation',
        'Turkey'
    )
    education = (
        "Less than a Bachelors",
        "Bachelor’s degree",
        "Master’s degree",
        "Post grad",
    )

    country = st.selectbox("Country", countries)
    education = st.selectbox("Education Level", education)
    #hard to create an empty selectbox
    experience = st.slider("Years of Expeience", 0, 50, 0) #start, end,
and default value in the slider
```

```
    ok = st.button("Predict Salary")
    if ok:
        X = np.array([[country, education, experience ]])
        X[:, 0] = le_country.transform(X[:,0])
        X[:, 1] = le_education.transform(X[:,1])
        X = X.astype(float)

        salary = regressor.predict(X)

        st.subheader(f"The estimated salary is ${salary[0]:.2f}")
```

The code for explore_page.py is:

```
import streamlit as st
import pandas as pd
import matplotlib.pyplot as plt



def shorten_categories(categories, cutoff):
    categorical_map = {}
    for i in range(len(categories)):
        if categories.values[i] >= cutoff:
            categorical_map[categories.index[i]] = categories.index[i]
        else:
            categorical_map[categories.index[i]] = 'Other'
    return categorical_map


def clean_experience(x):
    if x ==  'More than 50 years':
        return 50
    if x == 'Less than 1 year':
        return 0.5
    return float(x)
```

```
def clean_education(x):
    if 'Bachelor’s degree' in x:
        return 'Bachelor’s degree'
    if 'Master’s degree' in x:
        return 'Master’s degree'
    if 'Professional degree' in x or 'Other doctoral' in x:
        return 'Post grad'
    else:
        return 'Less than a Bachelors'


@st.cache_data #for every run, this code will perform all data cleaning
steps we used, this code saves this data as cache
def load_data():
    df = pd.read_csv("data\survey_results_public.csv")
    df = df.rename({"ConvertedCompYearly": "Salary"}, axis = 1)
    df = df[df["Salary"].notnull()]
    df = df.dropna()
    df = df[df["Employment"] == "Employed, full-time"]
    df = df.drop("Employment", axis=1)
    country_map = shorten_categories(df.Country.value_counts(), 250)
    df['Country'] = df['Country'].map(country_map)
    df = df[df["Salary"] <= 300000]
    df = df[df["Salary"] >= 30000]
    df = df[df['Country'] != 'Other']
    df['YearsCodePro'] = df['YearsCodePro'].apply(clean_experience)
    df['EdLevel'] = df['EdLevel'].apply(clean_education)


    return df
df = load_data()
def show_explore_page():
    st.title("Explore Software Engineer Salaries")
    st.write(
        """
    ### Stack Overflow Developer Survey 2020
```

```
    """

    )

    data = df["Country"].value_counts()

    fig1, ax1 = plt.subplots()

    ax1.pie(data, labels=data.index, shadow=True, startangle=90)

    ax1.axis("equal")  # Equal aspect ratio ensures that pie is drawn as a
circle.

    st.write("""#### Number of Data from different countries""")
#Multiline string in markdown format

    st.pyplot(fig1)

    st.write(

        """#### Mean Salary Based On Country"""

    )

    data =
df.groupby(["Country"])["Salary"].mean().sort_values(ascending=True)

    st.bar_chart(data)

    st.write(

        """#### Mean Salary Based On Experience"""

)

    data =
df.groupby(["YearsCodePro"])["Salary"].mean().sort_values(ascending=True)

    st.line_chart(data)
```

And finally, a small snippet of the dataset:

	A	B	C	D	E
1	Country	EdLevel	YearsCode	Employment	Salary
2					
3	Canada			Employed, full-time	
4	United Kingdom of Great Britain and Northern Ireland	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	5	Employed, full-time	40205
5	Israel	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	17	Employed, full-time	215232
6	United States of America	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	3	Employed, full-time	
7	Germany	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)		Student, full-time	
8	India	Secondary school (e.g. American high school, German Realschu		Student, part-time	
9	India	Some college/university study without earning a degree		Not employed, but looking for work	
10	Netherlands	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	6	Employed, full-time	49056
11	Croatia	Some college/university study without earning a de	30	Independent contractor, freelancer, or self-employed	
12	United Kingdom of Great Britain and Northern Ireland	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	2	Employed, full-time	60307
13	United States of America	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	10	Employed, full-time;Independent contractor, freelan	194400
14	United States of America	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	5	Employed, full-time	65000
15	Australia	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	15	Employed, part-time	
16	United States of America	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	5	Employed, full-time;Independent contractor, freelan	110000
17	Russian Federation	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	4	Independent contractor, freelancer, or self-employed	
18	Czech Republic	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	4	Employed, full-time	19224
19	Austria	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	10	Employed, full-time	202623
20	Austria	Secondary school (e.g. American high school, Germ	22	Employed, full-time	51192
21	Serbia	Some college/university study without earning a degree		Student, full-time	
22	Austria	Some college/university study without earning a degree		Student, part-time;Employed, part-time	
23	Italy	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	4	Employed, full-time	34126
24	Canada	Something else	20	Employed, full-time	97605
25	Netherlands	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	6	Employed, full-time	
26	Israel	Some college/university study without earning a de	40	Independent contractor, freelancer, or self-employed	
27	Germany	Masterâ€™s degree (M.A., M.S., M.Eng., MBA, etc.)	9	Employed, full-time	90647
28	United States of America	Bachelorâ€™s degree (B.A., B.S., B.Eng., etc.)	5	Employed, full-time	106960