**Project Ideas**

Before proceeding for the project, I came up with the following ideas to extract on. They will consist of ideas which I have never worked on, will have some familiarity to Data Science career and responses in daily life.

First of all, I searched on the biggest data center Kaggle for the hottest topic around the world and then gone through the datasets which I have not gone through.

So, the following are the topics/ ideas which I would love to be working on.

1. Sports

* Tennis
* Soccer
* Basketball

1. Data Science

* Career
* Resumes
* Types of Positions

We could try to predict the salary in the field of Data Science with respect to job title, company, geography.

**What data we need?**

Data collection are totally different for different projects and sometimes it becomes a cause of frustration. I personally read the job description on LinkedIn, Glassdoor also with the salary data.

**Data Scraper**

I personally got more data on Glassdoor as compare to LinkedIn. So I googled the scraper for the same. As a result, I got some Spyder code to scrape the data.

Links I used to get the data:

<https://towardsdatascience.com/selenium-tutorial-scraping-glassdoor-com-in-10-minutes-3d0915c6d905>

<https://github.com/arapfaik/scraping-glassdoor-selenium/blob/master/glassdoor%20scraping.ipynb>

**Data Cleaning**

As the data scraping tool worked, we will now clean the data with our purposes. Data cleaning is essential as the data scraped might have inappropriate data like no salary description, less job descriptions, etc. We scraped Glassdoor and LinkedIn data on the profession of Data Scientist. So we will now save the data in .csv/.xlsx file to further.

We will do the following:

1. Salary parsing
2. Company name text only
3. State Field
4. Age of company
5. Parsing of job description (python, etc)

**Exploratory Data Analysis**

The following points were recognized while performing the data analysis of the Cleaned Salary Dataset.

1. Job title & seniority
2. Fix state LA
3. Job description length
4. Competitor count
5. Hourly wage

**Model Building Process**

1. Choose relevant columns
2. Get dummy data
3. Train test split
4. Linear regression model
5. Lasso regression
6. Random forest
7. Tune models using GridSearchCV
8. Test ensembles

**Model Production**

We will deploy the model into production level and will use a web application using Python Streamlit library.