# **Portfolio content**

Website Link: https://gauravshetty98.github.io/portfolio/

## **Home**

**🔢 Data is more than numbers—it’s a roadmap to the future.** With expertise in AI, statistics, and predictive modeling, I develop systems that analyze, optimize, and forecast, helping businesses and researchers stay ahead of the curve.

### **Education:**

Rutgers University

Masters in Data Science

May 2025

Built expertise in machine learning, statistical modeling, and data mining, focusing on AI, big data, time series predictions, and GIS applications.

### **Projects:**

Building intelligent systems is more than just a profession—it's my passion. Whether it’s NLP, computer vision, or statistical modeling, I enjoy building intelligent systems that push the boundaries of AI. Here are some of my recent projects.

**NextBuys.co - A recommendation system**

Created a website that recommends products using a combination of similarity metrics and association rule mining.

**A Study on Adversarial Attacks and Defenses in Deep Learning Models**

This project examines how machine learning models can be vulnerable to adversarial attacks, exploring both attack strategies and defense mechanisms to enhance model security.

### **Experience:**

In the world of data, I’ve been the architect, the analyst, and the problem solver. I’ve tackled everything from predictive models to system designs, all with the goal of making things run better. Here's a look at what I've built so far.

#### **PGW:**

Statistics and economics intern

May 2024 - Dec 2024

As the sole Data Science intern of the customer programs department, my roles primarily involved developing predictive models and conducting statistical analyses on PGW’s customer data. My responsibilities spanned forecasting energy burdens, monthly customer counts, evaluating the impact of campaigns, and informing data-driven communication strategies. I delivered numerous projects, each involving collaboration with cross-functional teams, understanding their requirements, gathering useful data, and providing actionable insights and recommendations to stakeholders.

#### **Reliance Jio:**

Solutions Architect

Nov 2020 - Jul 2023

As a key member of the Design and Architecture team, I played an integral role in developing and delivering scalable telecom solutions. From creating systems for online SIM ordering to integrating new telecom regions, I ensured designs met business needs and were smoothly implemented. I also streamlined team workflows with custom Python tools to improve efficiency and reduce problem-solving time.

### **Publications:**

Curious about my research? Explore my published papers and see how data turns into discovery:

Sentiment Analysis and Classification on Twitter Spam Account Dataset

Developed a sentiment analysis model using VADER and a Random Forest classifier to accurately identify Twitter spam accounts, achieving 94% accuracy.

Type-II Diabetes detection using Decision-tree based Ensemble of Classifiers

Investigated the use of Decision Tree-based Ensemble classifiers for Type-II diabetes detection, achieving a maximum accuracy of 93.5% with XGBoost.

## 

## 

## 

## Experience Page:

### **Philadelphia Gas Works:**

### Statistics and Economics Intern

### As the sole Data Science intern of the customer programs department, my roles primarily involved developing predictive models and conducting statistical analyses on PGW’s customer data. My responsibilities spanned forecasting energy burdens, monthly customer counts, evaluating the impact of campaigns, and informing data-driven communication strategies. I delivered numerous projects, each involving collaboration with cross-functional teams, understanding their requirements, gathering useful data, and providing actionable insights and recommendations to stakeholders. Some of my highest-impact projects completed at PGW are elaborated below.

### 

### **1. Monthly Customer Enrollment Forecasting Using Time-Series Analysis**

* **Objective**: To develop a model to help with budgeting and resource allocation by forecasting the monthly number of customers enrolling in specific programs.
* **Business Impact**: The model provided a reliable way to predict monthly enrollments, enabling the organization to plan budgets, allocate resources more effectively, and optimize programs. By accounting for seasonality and external factors, it significantly improved the accuracy of quarterly forecasts and supported smarter decision-making.
* **Methodology**: I started with exploratory data analysis (EDA) in Python to identify trends, seasonal patterns, and external factors influencing enrollments. After cleaning and preprocessing the data, I used the Prophet time series model to capture seasonality, achieving a mean absolute percentage error (MAPE) of less than 10%. I fine-tuned the model iteratively by adjusting hyperparameters and incorporating external regressors for improved accuracy.
* **Tools**: I cleaned and preprocessed the data using Python and SQL, then developed the model in Python. The final results were delivered through an interactive Excel dashboard, providing easy access and usability.
* **Challenges**: The most challenging aspect of the project was building the model. I addressed this by conducting extensive research, experimenting with various models, and fine-tuning them iteratively to optimize performance and ensure the best results.
* **Results**: The model was well-received by the team for its accuracy, accessibility, and ability to incorporate external factors into projections. These projections were reviewed during monthly meetings to evaluate customer growth and optimize budget allocation for discount programs.

### 

### **2. Analysis of Low-Income Customer Behavior**

* **Objective**: The company wanted to analyze the effectiveness and participation rates of various financial assistance programs and identify the major factors affecting these participation rates.
* **Business Impact**: The analysis helped the organization gain a deeper understanding of its low-income customer base and address critical business questions, such as identifying factors affecting participation rates, evaluating the effectiveness of existing communication strategies, uncovering barriers to program engagement, and determining correlations between demographic or behavioral patterns and program participation.
* **Methodology**: I conducted exploratory data analysis (EDA) on a dataset of over 50,000 customers, utilizing pivot tables, scatter plots, histograms, and data grouping to uncover patterns and trends. Correlation analysis was performed to identify key relationships between demographic factors and program participation rates. The data was cleaned and preprocessed in Python, with insights visualized in Tableau.
* **Tools**: I utilized Python and SQL for data cleaning and preprocessing, with analysis and visualizations created in Tableau.
* **Challenges**: Cleaning and merging data from multiple sources was a key challenge, which I effectively addressed using Python's pandas library.
* **Results**: The analysis revealed factors, such as age and assistance type, that affected participation rates. It also raised new questions about whether current outreach methods were effective, forming the basis for a survey conducted among customers to explore similar questions.

### 

### **3. Integrating City Government Datasets to Identify Potential Candidates for Financial Assistance Programs**

* **Objective**: As an intern, one of my tasks was to research publicly available datasets that could help identify potential customers for financial assistance programs.
* **Business Impact**: This initiative ensured early recognition of low-income customers, enabling better financial aid distribution, reducing payment delinquencies, and optimizing the utilization of government-supported program budgets.
* **Methodology**: I researched publicly available datasets from sources like the [Census](https://data.census.gov/) and [OpenDataPhilly](https://opendataphilly.org/). ~~Using these sources, I identified low-income housing locations in the city and integrated this data with our dataset to pinpoint potential customers.~~  
  I researched publicly available datasets like the U.S. Census and OpenDataPhilly, focusing on socio-economic indicators and low-income housing. Using Python, I preprocessed and standardized the data, performed EDA to map housing locations, and integrated this with our customer data. Geospatial analysis and geocoding were applied to ensure accurate mapping for targeted outreach.
* **Tools**: I preprocessed the data using Python to standardize and align it with the company’s formatting before merging it with our customer data.
* **Challenges**: The main challenge was identifying publicly available datasets that addressed key organizational issues while ensuring data quality. I tackled this by prioritizing the organization's challenges, researching datasets relevant to the top problems, and focusing on government-provided or well-documented data to maintain quality standards.
* **Results**: I delivered a list of potential leads for low-income programs, enabling the team to execute targeted outreach campaigns and inform customers about financial assistance programs they might qualify for.

### 

### 

### **4. Energy Burden Forecasting and Analysis**

* **Objective**: The purpose of this project was to identify disparities in energy usage and energy burden between customer programs, forecast future burdens, and provide actionable insights through a user-friendly dashboard.
* **Business Impact**: This tool provided a deeper understanding of customer energy usage, helping the organization reduce operational inefficiencies, enhance program performance, and better support customers with high energy burdens.
* **Methodology/Tools**: I integrated city-provided [property assessment data](https://opendataphilly.org/datasets/philadelphia-properties-and-assessment-history/) with customer energy usage records using SQL, enabling an analysis of energy consumption alongside housing attributes such as construction year, living area, and building conditions.  
  I performed data integration by merging city-provided property assessment data with customer energy usage records using SQL, leveraging joins and aggregation techniques to align the datasets. I applied feature engineering to derive meaningful attributes from the housing data, such as energy efficiency indicators based on construction year and building conditions. For advanced analysis, I used time-series analysis techniques to identify trends in energy consumption, applying seasonal decomposition to account for cyclical patterns. I developed a highly interactive Tableau dashboard using calculated fields and parameters to visualize energy usage distributions, enabling real-time comparisons across customer programs.
* **Methodology/Tools**: I developed an interactive Tableau dashboard to visualize energy usage trends across customer programs. The dashboard provided detailed comparisons based on housing attributes like construction year and living area. Additionally, I incorporated a feature to estimate energy burden using historical averages, total customer counts, and customizable group definitions for tailored analysis.
* **Challenges**: The biggest challenge was creating an interactive Tableau dashboard that directly addressed business needs while allowing users the flexibility to explore the data further. I overcame this by studying effective dashboard designs, collaborating closely with senior team members to understand stakeholder priorities, and organizing the dashboard into clear, functional sections. Using SQL and Tableau, I incorporated essential features and ensured the tool was intuitive and impactful.  
  ​​The project faced several key challenges, starting with the high dimensionality of the dataset, which contained over 50 features. Managing and analyzing such a large number of variables was complex, requiring feature selection techniques and dimensionality reduction to ensure meaningful insights. Integrating data from multiple sources, including city-provided property assessments and internal energy usage records, involved extensive data wrangling to ensure compatibility and handle inconsistencies. Missing data and outliers posed additional difficulties, which I addressed through imputation techniques and outlier detection methods. Additionally, designing a user-friendly and interactive Tableau dashboard that could handle detailed visualizations while remaining intuitive for users required iterative development and close collaboration with stakeholders.
* **Results**: I created a comprehensive Tableau dashboard that offered insights into energy usage patterns, facilitated comparisons between customer programs, and provided forecasting capabilities for energy burden, significantly improving program analysis and decision-making.

### 

### **6. Survey to Explore Financial Assistance Barriers**

* **Objective**: Previous analysis revealed age as a key factor affecting participation rates in financial assistance programs. To further explore this and address broader questions surrounding low participation rates, I conducted a survey to assess program awareness, effectiveness, and the impact of various factors on engagement.
* **Business Impact**: The survey provided critical insights into customer perceptions of financial assistance programs, the effectiveness of existing outreach campaigns, and areas for improvement. By addressing key business questions, it enabled the organization to refine awareness efforts, enhance program engagement, and optimize communication strategies to better meet customer needs.
* **Methodology**: I spearheaded the survey with the help of [Bellomy Research](https://www.bellomy.com/) among 900 customers. The survey posed questions related to program awareness, preferred communication methods, customer satisfaction, and other relevant topics.
* **Tools**: I merged the survey responses with customer data using Excel and derived analytical insights using Tableau and Python.
* **Challenges**: Formatting the survey responses for analysis was a significant hurdle, particularly because much of the data was free text. I addressed this by using Excel and Python to preprocess and organize the responses into a structured format, ensuring they were ready for analysis.
* **Results**: The analysis identified ineffective communication channels and highlighted preferred outreach methods for different customer groups. These findings enabled the organization to refine awareness campaigns, introduce new outreach strategies, and improve overall program engagement and reach.

### 

### 

### **7. Analysing the effectiveness of our recertification strategy**

* **Objective:** The objective of this project was to analyze the annual recertification process required for customers to remain enrolled in low-income programs, identify stages with the highest drop-off rates, assess its effectiveness, and provide insights to help the organization strategize and forecast future customer retention.
* **Business Impact:** The analysis provided a deeper understanding of the recertification process, helping the organization reduce customer attrition, improve the efficiency of their programs, and better anticipate recertification volumes to allocate resources effectively.
* **Methodology/Tools:** Data preprocessing was done using SQL and data visualization was done using Tableau. Developed a dashboard to track customer retention through each stage, identify drop-off points, and visualize future recertifications.
* **Challenges:** The biggest challenge was visualizing the recertification process, as its 45-day timeline varied for each customer, making it difficult to showcase monthly drop-offs. Unclean data further complicated the task. I addressed this by performing data aggregation and grouping customers by their stage and remaining days in the process, enabling clear and interpretable visualizations that highlighted monthly drop-offs effectively.
* **Result:** The dashboard helped in identifying stages that faced the most drop-off thus improving the overall process, it also provided details about upcoming recertifications helping the organization in better planning.

### 

## Reliance Jio Infocom LTD

## Solutions Architect

As a key member of the Design and Architecture team, I played an integral role in developing and delivering scalable telecom solutions. From creating systems for online SIM ordering to integrating new telecom regions, I ensured designs met business needs and were smoothly implemented. I also streamlined team workflows with custom Python tools to improve efficiency and reduce problem-solving time.

### **1. On-Demand SIM Activation and Doorstep Delivery System**

* **Objective**: The company wanted a system that would allow users to order an activated SIM card directly from the company website or mobile app and have it delivered to their doorstep.
* **Business Impact**: This system aimed to increase the reach of our products and make it easier for customers to place orders. As no similar system was available in the market at the time, it gave our product a competitive advantage.
* **Methodology**: Researched competitor solutions to identify shortcomings and customer-preferred workflows, guiding improvements in system design. Collaborated with development teams to assess available APIs and system integrations. After extensive brainstorming, designed a secure, end-to-end system that adhered to data privacy and security guidelines. Integrated advanced computer vision models for fraud prevention using face detection. Presented the system architecture to over 500 stakeholders and actively oversaw development and deployment, providing continuous support to resolve issues and ensure a smooth rollout.
* **Tools**: I used **SOAP UI** and **Postman** to test APIs across existing systems, ensuring functionality and reliability. **SQL** was utilized to query databases and validate data flow, ensuring consistency and accuracy. For issue tracking and management, I leveraged **Azure DevOps**, streamlining debugging and collaboration across teams.
* **Challenges**: One of the main challenges was the scale of the project and the number of stakeholders involved. Designing a solution that met everyone’s requirements was difficult. We addressed this by holding regular meetings with different teams to provide constant updates on the system and gather feedback. Additionally, we ensured that every stakeholder understood the priority and importance of various requirements, facilitating a smooth design and development cycle.
* **Results**: The system was deployed in record time, ahead of any other telecom provider, making our company the first to offer such a service at the time. Check it out for yourself: https://www.jio.com/selfcare/selfkyc/

### **2. Central Coupon Management System**

* **Objective:** The organization ran multiple campaigns to onboard customers with various coupons and discount vouchers but lacked a central system to manage them, making it difficult to access historical data or use existing APIs.
* **Business Impact:** A centralized system would standardize campaign management, reducing complexity and ensuring consistency across all campaigns. It would also simplify coupon generation, management, and validation for on-field employees for all campaigns
* **Methodology:** I conducted a technical analysis of historical campaigns to identify commonalities in API structures, coupon generation mechanisms, and data handling. Leveraging existing APIs, I designed a scalable system that could manage past campaigns while seamlessly integrating new ones. Using **SOAP UI** and **Postman**, I tested and optimized API workflows to ensure smooth data exchange. **SQL** was employed for data validation, ensuring accuracy in coupon tracking and campaign records. The system was built with modularity in mind, enabling flexibility for future campaign integrations while maintaining consistency and security in coupon management.
* **Tools:** I usedSoap and Postman to test and design APIs. SQL was used to do data validation and testing.
* **Challenges:** The biggest challenge was analyzing historical campaigns and consolidating diverse coupon generation systems into a unified framework. Each campaign had its own technical flow, APIs, and validation mechanisms, requiring extensive research to identify commonalities and standardize processes. Additionally, collaborating with multiple teams—including marketing, campaign management, and development—was complex, as each had different requirements and priorities. Balancing these inputs while ensuring system compatibility and scalability required careful negotiation, iterative refinement, and rigorous API testing. By thoroughly reviewing documentation, consulting senior members, and systematically testing integrations, I was able to develop a cohesive solution that met all business and technical needs.
* **Result:** The result of this project was the successful development of a centralized campaign management system that streamlined coupon generation, validation, and tracking across all customer acquisition campaigns. The system standardized workflows, reducing operational complexity and ensuring seamless integration with existing APIs. It enabled marketing and on-field teams to efficiently manage and retrieve historical campaign data, improving decision-making and reducing redundancy in new campaigns. The system became a foundational tool for future campaigns, significantly improving efficiency and scalability in promotional efforts.

### **3. Python Programs for Operational Efficiency**

* **Objective:** I identified several areas within the company that could be improved through Python automation, such as reducing penalties from faulty invoices and addressing data extraction issues while fixing bugs. I developed Python-based tools to streamline these processes and effectively resolve these challenges.
* **Business Impact:** The Python-based programs significantly enhanced efficiency by automating data extraction from invoices and PDFs, reducing manual effort and analysis time. The software also identified defective invoices, minimizing penalties from external audits and improving compliance.
* **Methodology:** The approach focused on identifying repetitive tasks that consumed significant time for the team and determining whether a software solution could perform them more efficiently. After developing the tools, I presented them to the team, provided thorough documentation on their usage, and included guidelines for future modifications.
* **Tools:** The tools were developed in Python, with Azure DevOps used for version control. Since the primary focus was working with PDF - extracting data and storing it in a structured format, libraries such as OpenCV, tabula-py, and PyMuPDF were utilized.
* **Challenges**: //
* **Result:** The program detected 90% of defective bills, significantly reducing penalty costs for the company. It also enabled efficient extraction of text and image data from invoice PDFs, cutting analysis time for the testing team.

### **4. Youth Customer Onboarding and JioPhone Sales Integration**

* **Objective:** The organization wanted to improve the accessibility of our services and our products like JioPhone by introducing them to younger demographics. The current system didn’t have proper checks in place to cater to such needs.
* **Business Impact:** Since JioPhone was a product that was extensively used in rural areas, such a system would increase the reach of an already popular product.
* **Methodology:** I integrated the existing system with image recognition technology using [Hyperverge](https://hyperverge.co/) to authenticate younger customers through valid identity documents. Additionally, I maintained and updated the JioPhone sales system to ensure it remained aligned with our other workflows.
* **Tools:** I primarily used Visio and UML to visualize the system architecture and data flow. For data validation and testing, I utilized SQL, Postman, and SOAP.
* **Challenges:** Understanding the JioPhone system was challenging due to its complexity and age. To overcome this, I thoroughly reviewed the available documentation, sought guidance from senior team members, and conducted extensive testing of the system and its API calls. This hands-on approach helped me gain a comprehensive understanding of the system's logic and functionality.
* **Results:** The system and the product are still an integral part of the company revenue and is one of the crucial products, especially popular among aged and rural customers.
* URL: https://www.jio.com/en-in/jiophone

### 

# 

## Projects:

## 

## NextBuys.co

**NextBuys.co is a** recommendation engine system that enhances product recommendations using Amazon transaction data. Leveraging a dataset of over 1.8 million transactions across 1,800 product categories, the system preprocesses data to identify multi-item purchases and applies cosine similarity to refine product categorization. Algorithms like FP-Growth and metrics like cosine similarity have been used to optimize these recommendations. A customer feedback loop to change recommendations based on user input.

**Objective:** To build a recommendation system capable of recommending items based on purchase trends and also based on similarity.

**Methodology:** Utilized FP-Growth algorithm to identify items that are frequently bought together and used metrics such as lift and confidence to refine these suggestions. Implemented cosine similarity to identify items that are similar to each other. Also designed a customer feedback loop to let the users like or dislike the recommendations based on which they are altered.

**Tools:** Recommendations and dataset was created using python and stored in SQL. Webpage was designed using HTML, CSS and Javascript.

**Outcome:** Was able to develop an end-to end system combining ML and Web-dev techniques.

## A Study on Adversarial Attacks and Defenses in Deep Learning Models

This project explores **inference attacks** on machine learning models, highlighting both **attack strategies** and **defense mechanisms**. As machine learning becomes increasingly prevalent, it is crucial to recognize that these systems are also vulnerable to adversarial attacks. Using a **PyTorch-based computer vision model** for vehicle damage assessment, we demonstrate how an attacker can exploit weaknesses using the **Fast Gradient Sign Method (FGSM)** to significantly degrade model accuracy. Additionally, we showcase **defensive techniques** such as **Gaussian Noise Augmentation and adversarial training** to mitigate these attacks and restore model performance. This project underscores the importance of **robust AI security** in real-world applications.

**Objective:** This project aims to analyze the impact of adversarial attacks on deep learning models, focusing on inference attacks and defense mechanisms in computer vision applications. The objective is to highlight the vulnerabilities of machine learning models to adversarial manipulation and evaluate countermeasures to ensure secure and reliable AI systems in real-world applications.

**Methodology:** The project follows a structured approach, beginning with the development and training of a computer vision model for vehicle damage assessment using PyTorch. After establishing a baseline accuracy of 77%, we implement FGSM attacks to craft adversarial samples and assess their impact on model performance. The model performance dropped to 16% based on these attacks. We also look into defensive techniques such as Gaussian Noise Augmentation, introducing controlled noise during training, and implement adversarial training, which retrains the model with adversarial examples to improve its resistance. The effectiveness of these defenses is measured by evaluating key performance metrics, including accuracy degradation, robustness improvements, and adversarial resistance. Visualization techniques such as confusion matrices and perturbation heatmaps further illustrate the attack and defense dynamics.

**Tools:** Utilized the ART library in python to program the attacks. PyTorch and Torchvision to build the convolutional neural network and LIME for model explainability.

**Outcome:** This study demonstrates how even machine learning models are exposed to security risks like these attacks which can degrade the model performance. It also shows ways to defend against these attack emphasizing the necessity of adversarial defense mechanisms in securing AI applications.

## Assessing Heart Failure Risk: Statistical Modeling with Logistic Regression

​​This project explores the use of logistic regression for predicting heart failure survival, leveraging statistical modeling to understand key health predictors. Exploratory data analysis (EDA) was performed to uncover patterns and relationships within patient records. Techniques like stepwise selection were used to optimize model performance while maintaining interpretability. Model validity was assessed through residual analysis, while odds ratios provided insights into the impact of various health factors on survival outcomes. This study highlights the role of data-driven decision-making in healthcare analytics.

**Objective:** This project aims to assess heart failure risk by leveraging **logistic regression** to predict patient survival based on key health indicators. It also explores the different statistical methods used for model optimization, feature selection and model explainability.

**Methodology:** The study begins with **exploratory data analysis (EDA)** to identify patterns, correlations, and outliers within patient records. Feature selection is performed using **stepwise regression**, ensuring that only the most relevant predictors contribute to the logistic regression model. Model accuracy and reliability are assessed through **residual analysis and validation metrics** such as chi square tests and confusion matrices. Additionally, **odds ratios** are computed to quantify the impact of individual health factors on survival probability, providing a clear interpretation of the model’s predictions.

**Tools:** Used R as the main programming tool to conduct EDA. Using generalized linear models to build the logistic regression model.

**Outcomes:** The project shows in the emerging world of complex neural networks, how a statistical model can provide accurate results while maintaining the simplicity and being interpretable. It also showed key risk factors such as age, ejection fraction, and creatinine levels significantly influence survival probability, as evidenced by their odds ratios.

## Deep Learning for Sentiment Analysis: IMDb Movie Reviews with LSTMs

This project explores **sentiment analysis** on a **movie reviews dataset** using **bi-directional Long Short-Term Memory (LSTM) networks**, showcasing the application of **deep learning** in **Natural Language Processing (NLP)**. The model is designed to analyze the sentiment of movie reviews, leveraging **one-hot encoding** for vectorization to capture contextual dependencies in text data. A **sequential neural network** with multiple LSTM layers and a dense output layer using a **sigmoid activation function** was implemented for classification. The project examines the network architecture and its performance, highlighting the effectiveness of **recurrent neural networks (RNNs)** in sentiment analysis.

**Objective:** The aim of this project was to explore the applications of deep learning algorithms like Long Short-Term Memory networks in the field of Natural Language Procesing. The study demonstrates the use of RNNs and their performance in classifying whether a particular moview review was positive or negative.

**Methodology:** The dataset is preprocessed through one-hot encoding to transform text into numerical vectors, allowing the LSTM model to effectively capture word dependencies. A bi-directional LSTM architecture is implemented with multiple layers followed by a sigmoid activation function for binary classification. The model is trained and evaluated using standard metrics such as accuracy, precision, recall, and F1-score, while techniques like dropout regularization are applied to prevent overfitting.

**Tools:** The program was coded in Python. Libraries like Tensorflow, Keras and scikit-learn were used for model building and evaluation.

**Outcomes:** The study underscores the importance of deep learning in NLP applications, particularly in understanding sentiment and context within textual data. The findings suggest that LSTMs are well-suited for sentiment analysis, paving the way for further improvements using attention mechanisms or transformer-based models.

## ML-Driven Industry Analytics: LinkedIn, BERT, and GIS in Action

This project aimed to provide industry analytics based on a user's LinkedIn profile, offering insights into salary trends, job openings, and industry shifts pre- and post-COVID. A dashboard built with Plotly Dash integrated the LinkedIn API to fetch user data and analyzed Bureau of Labor Statistics datasets. Vectorization techniques and cosine similarity were used to identify similar industries, enhanced by Hugging Face’s BERT model for improved embeddings. GeoPandas helped in visualising geographic trends for those industries. The result? A data-driven, interactive tool that delivers personalized insights into industry projections, leveraging data science techniques and public datasets to inform career decisions.

**Objective:** The project aimed to leverage machine learning and NLP techniques to deliver personalized career insights by analyzing data from the Bureau of Labor Statistics and the LinkedIn API. The goal is to offer insights into salary trends, job openings, and industry shifts before and after COVID-19.

**Methodology:** The project utilizes a Plotly Dash-based interactive dashboard to visualize the analysis. User details fetched from Linkedin API are preprocessed before applying NLP techniques to find user’s industry data from Bureau Of Labour statistics. A combination of word embeddings derived through BERT and cosine similarity metric is used to identify the user’s industry from the BLS database. In-depth analysis is done on this data to showcase different kinds of trends in industry including geographic hotspots for the industry with the help of GIS libraries like GeoPandas and Folium.

**Tools:** Linkedin API was used to fetch user details in real time. Libraries like BERT and Cosine similarity were used for text processing. Visualization libraries like Plotly and GIS libraries like GeoPandas, Folium were also utilized. The website was created using Plotly Dash.

**Outcomes:** The study helped in creating a system for providing career insights based on a users industry and displayed how different Data Science techniques can be used in real world applications. It highlights how these techniques can be used in transforming labor market insights into actionable career decisions. This project lays the foundation for AI-powered career advisory tools, demonstrating the potential of data science in workforce analytics.

NBA Player Performance Analysis

The aim of this project

## Scrapyard: Used [property assessment data](https://opendataphilly.org/datasets/philadelphia-properties-and-assessment-history/) provided by the city and integrated it with customer energy usage data using Tableau and SQL. This provided a way to analyse energy usage along with housing attributes like year built, exterior conditions, living area,etc.

* Developed an interactive dashboard in Tableau visualizing energy usage patterns among customers in different programs living under similar conditions. Also added a feature to estimate energy burden based on total customer count and historical means.
* It could answer the question of whether fixed bill customers are using more energy than regular customers while living in similar conditions. It would provide the company with an easy access tool to view energy usage historical patterns among different groups, a way to view housing attributes of customers belonging to different programs, comparing the energy usage between different programs and also a way to estimate the energy burden using historical averages based on custom customers groups that a user could create.
* One of the questions the organization had was how the energy usage differed between different programs, if there are programs where customers use more energy and also a way to estimate the energy burden.
* The organization wanted to compare the energy burden between our fixed bill customers and regular customers. They also wanted an energy burden forecasting model for the different customer programs.
* Utilized publicly available property details from the city government to compare living conditions and energy usage among customers from different programs.
* Developed an interactive Tableau dashboard to visualize how the energy usage differed based on customer groups and living conditions proving customers with fixed monthly bills were using more energy.
* Delivered an interactive dashboard that enabled the organization to analyze energy usage and burden across customer groups, compare fixed bill and regular customers under similar living conditions, and explore housing attributes of different programs. The tool provided actionable insights into historical energy patterns, allowed custom group analyses, and offered an energy burden estimation feature, helping in decision making.
* Dashboard also included a section to forecast monthly energy burden using historic averages and customer estimates
* The analysis provided insights on communication channels which were ineffective, it provided information on new outreach methods and helping in answering various business questions.
* Anlaysis showed how age was a factor in participation rates among financial assistance programs. A survey would help in answering some of the business questions surrounding participation rates and the factors effecting it.
* Spearheaded a survey among 900 customers which looked into answering questions related to low participation rate in financial assistance programs among some customer groups
* Designed and executed a survey targeting 900 customers. Survey consisted of questions ranging from awareness of programs to effectiveness of those programs.
* Analyzed survey responses to analyze reasons behind reduced participation. Validated customer responses by merging survey responses with company data.
* The survey provided insights into the preferred communication mediums of different customer groups and identified ineffective channels. This enabled us to enhance existing outreach methods and introduce new ones, improving overall engagement and reach
* .The program was able to detect 90% of the defective bills thus reducing the company cost on penalties. It also provided our team with a way to extract textual and Image data from invoice PDFs, which reduced our analysis time when testing bulk orders.
* The tools made some of the processes more efficient by reducing time to extract data from invoices and pdf. It also helped in reducing the penalty cost for faulty documents that were being found by the audit teams.
* The audit team used to face a lot of issues with defective bills and invoices of customers. The organization was also penalized whenever a defective invoice was detected during external audits. The team wanted a simple way to analyze these documents and also a way to extract the details present in it.
* Developed a python program which helped with running checks in the documents to reduce manual work of analysing basic details of the customers.
* The program also extracted text and image data from invoices, cleaning it, and storing it in an accessible format. This function was primarily helpful to the testing team who needed a tool to validate final data during testing runs.
* Similarly, also created a tool for data extraction from Business Intelligence (BI) and Revenue Assurance Fraud Management (RAFM) reports, reducing analysis time by 50%.
* These tools helped the team and organization by reducing the time taken for testing and analysis. It was successful in catching 90% of the defective bills thus also reducing the penalties for the same.
* The major challenge in this project was research through historical campaigns and choose which APIs would be better in the system. Each campaign had their own technical flow and APIs to generate coupons. Understanding all the different services and how to use them was difficult. The way I resolved this was by spending time to read through the API documentations prepared by the team, asking for guidance from senior members whenever required and testing the different APIs and how the data flowed through our system. This helped in then building the final solution which fulfilled all the requirement. The system would make it easier to launch new campaigns and would make all new campaigns standardized following the same format. It would help in accessing historical data for analysis.  The organization used to run several campaigns where they used to onboard large groups of customers by providing them different coupons. There was no central system to manage and organize these different campaigns thus making it extremely difficult when starting with a new campaign or to check the historical data of an old campaign.
* Designed a central system using already developed APIs which had the capability to generate and validate coupon details based on the campaign.
* Made campaign management less complex, easier to generate and validate coupons for on-field employees and reduced the complexity of starting new campaigns
* Customers in our low-income programs were required to recertify annually to remain enrolled. This process led to many customers being dropped from the program due to various reasons. The primary objective of this analysis was to The primary objective was to analyze the recertification process by answering key questions, such as how effective it was, when customers typically recertified, and at which stage of the process the most drop-offs occurred.
* Used SQL for data preprocessing and Tableau dashboard to visualize the total number of customers present and dropping after each stage.
* Dashboard also helped in displaying future recertifications helping the company to strategize accordingly.
* The dashboard showed which stages were most effective and which stages faced the most drop outs. The predicted recertification numbers also helped in forecasting how many customers might drop off in the coming months.

Created a dashboard using plotly dash where a user can look at industry analytics using linkedin handle. Integrated LinkedIn API to fetch user data. The dashboard included different graphs ranging from mean salary change pre and post covid. Change in annual openings pre-covid and post covid. Industry data was sourced through Beaurau of Labour Statistics. The analysis used to identify similar industries based on your linkedin profile using similarity metrics like cosine similarity and vectorization techniques. Used huggingface BERT model to create embedding vectors which improved the similarity results. Also used GIS libraries like geoPandas to show areas of interest for different industries. The objective of this product was to show it to a user a brief look into their respective industries and how the projections are compared to pre covid and post covid using publicly available data and data science techniques.