Jesús Cantú Jr.

Chicago, IL | (956) 533-9738 | jesus.cantu217@gmail.com | LinkedIn | GitHub | Medium

Data Scientist & ML Engineer

With over a decade of interdisciplinary experience in data analytics spanning Computer Science, Sociology, and Public Health, I bring a diverse range of skills to the forefront. Proficient in Python, SQL, and R, I excel in machine learning and data science tools such as TensorFlow, PyTorch, and the Apache Hadoop Ecosystem. Wearing multiple hats, I adeptly navigate between machine learning model development, data processing, and hardware-software integration, crucial in contributing to successful multidisciplinary projects. Demonstrating leadership, I have managed a team of 5 researchers, resulting in key publications on infectious disease seasonality. My recent GitHub projects highlight my ability to analyze data effectively and drive societal impact, while my Medium articles aim to simplify complex technical concepts. Committed to community engagement, I leverage my technical skills and interdisciplinary knowledge to tackle real-world challenges.

Education

- **M.S. in Software Engineering – Loyola University, Chicago | May 2023**
- **B.A. in Sociology – Princeton University | June 2017**

Technical Skills

- **Programming Languages:** Python, SQL, R, C#, Java, Bash, HTML
- **Data Science & ML Tools:** TensorFlow, PyTorch, Scikit-Learn, Pandas, NumPy, Matplotlib, Seaborn, ApacheAirflow, Apache Hadoop Ecosystem (HBase, Phoenix, Hive, Impala, Spark)
- **Data Analysis:** Unsupervised Learning (K-Means, Clustering, PCA), Supervised Learning (Linear & Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, K-NN, XGBoost)
- **Visualization Tools:** Shiny, Tableau, Power BI
- **Databases:** MySQL, PostgreSQL, SQLAlchemy, MongoDB
- **Cloud Services:** AWS, GCP, Azure, Databricks
- **DevOps:** Docker, Git, MLflow

Experience

TMW Center for Early Learning + Public Health **Oct 2023 – Current**
Applied Data Fellow

- Spearhead the development and validation of advanced machine learning algorithms for classification, leveraging Python, the Hugging Face Transformers library, and open-source models (e.g., wav2vec2 and whisperX) to analyze conversational dynamics between child-caregiver pairs using large audio datasets (> 5GB).
- Lead the design and implementation of robust data engineering pipelines, ensuring data reproducibility and scalability on AWS infrastructure. Utilize Python scripting and AWS tools to manage and process vast datasets efficiently.
- Play a key role in the design and integration of a wearable device, collaborating closely with hardware and firmware teams to ensure seamless interaction between device components and machine learning algorithms.

MEDIUM **June 2023 – Current**
Technical Writer

- Authored and published over 20 comprehensive articles on pivotal topics within data science, machine learning, and data engineering, amassing a readership of over 2K views per month. Demonstrated a keen ability to translate complex technical concepts into accessible, engaging content.
- Contributed to the *LatinXinAI* publication with an article on the process of training and deploying Large Language Models (LLMs), with a focus on advanced techniques like transfer learning, attention mechanisms, and regularization.

LOYOLA UNIVERSITY – CHICAGO **May 2022 – May 2023**
Data Scientist, Dpt. Of Computer Science

- Partnered with interdisciplinary teams of psychologists at Arizona State University and computer scientists at Loyola University in the identification of new research areas. Developed and executed long-term research strategies. Submitted and won a Google Research Grant valued at ~$60k in 2023.
- Clearly defined, implemented, and evaluated the data structure and experimental framework necessary to construct new natural language processing (NLP) models. Actively contributed to data science pipelines via the creation of new datasets. Utilized Python to collect and analyze 15M+ social media data points through various APIs for cyberbullying/hatespeech detection.

INVERSA LEATHERS **July 2021 – Feb 2022**
Data Scientist

- Transformed business questions into research projects with the appropriate methods and tools. Directed independent research studies that estimated the geographic spread of lionfish through spatial analysis of GEBCO 2020 bathymetry data in R and QGIS. Enabled the team to raise $2M from the private market after assisting in identifying the total addressable market of a scaled response to the invasive species.
- Assisted in the development of commercial fishing strategies by developing testable hypotheses that evaluated KPIs. Conducted statistical analyses, forecasted results, and aided the research and development team in executing well-grounded recommendations to the executive leadership team and other stakeholders.

UNIVERSITY OF CHICAGO **Oct 2019 – Jun 2021**
Research Scientist, Dpt. Of Ecology and Evolution

- Planned, directed, and conducted experimental research utilizing mechanistic SIR (Susceptible; Infected; Recovered) models alongside data analytics in R and translated research findings into targeted health interventions.
- Simulated the effects of temperature, humidity, and rainfall on the seasonality of malaria transmission. Wrangled and visualized time series data (~500K+ observations). Compiled the results in a comprehensive report.

COLUMBIA UNIVERSITY **Feb 2018 – Sep 2019**
Lab Manager & Research Associate, Dpt. Of Environmental Health Sciences

- Supervised research program operations and advised a team of 5 researchers.
- Built and simulated a dynamic transmission model for chickenpox using R and C++ snippets. Fitted the model, a partially observed Markov process, using 80+ years of time series, epidemiological and demographic event data.
- Collected and utilized 1st/3rd party healthcare data (~25K+ observations) to measure varicella vaccine efficacy. Presented study findings at an international ecology and infectious disease meeting.
- Drafted and reviewed research papers for publication.

PRINCETON UNIVERSITY **Sep 2016 – May 2017**
Researcher & Data Analyst, Dpt. Of Sociology

- Used exploratory data analysis, inferential statistics (t-tests, correlation analysis, ANOVA), and predictive analytics (multivariate and random-effects regression models) in R to study the factors contributing to the high prevalence of teenage births in Texas and prescribe better social/health policies.
- Collected, cleaned, wrangled, and analyzed data from 2005 to 2014 relating to 8 different social determinants of teenage pregnancy across 254 different counties in Texas (~10K observations).

UNIVERSITY OF GEORGIA **Summer 2016**
Research Intern, Odum School of Ecology

- Discovered that unreported infections by asymptomatic individuals contribute minimally to the spread of the Hepatitis A virus across populations.
- Simulated outbreaks of Hepatitis A using R, focusing on different vaccination strategies and estimated the number of asymptomatic Hepatitis A cases over time with the aim of identifying their role in sustaining transmission.

PRINCETON UNIVERSITY **Nov 2015 – May 2016**

Research Assistant, Dpt. Of Ecology and Evolution

- Used descriptive statistics and conducted multivariate regression analysis in R to study the socio-ecological factors contributing to the high incidence of varicella in Texas, particularly among border counties.
- Prepared and analyzed annual, county-level data for varicella cases and immunization coverage among children (5-12 years old), as well as population and demographic information for the State of Texas from 2005 to 2014.
- Presented study findings at an international ecology and infectious disease meeting.

UNIVERSITY OF COLORADO SCHOOL OF MEDICINE **Summer 2014**

Research Intern, Webb-Waring Center

- Conducted deep-dive research on the positive selection factors for the MUC5B variant (rs35705950) and produced a literature review of idiopathic pulmonary fibrosis (IPF).
- Identified a possible mechanism through which MUC5B could convey a significant fitness advantage and proposed a potential agent of directional selection that may contribute to the current prevalence and worldwide distribution of the variant across ethnic populations, in relation to that of IPF.

Awards

- **Margaret H. Hamilton, High Achievement Award (2023):** Recognized as the top academic performer in the MS program across all computer science degrees at Loyola University based on GPA.
- **Gates Millennium Scholar (2013):** Distinguished as 1 of 1000 scholars nationwide to receive this esteemed scholarship.

Projects

Speech-to-Text Transcription and Analysis | GitHub | October 2023

- Tools & Technologies: Python, Google Cloud SDK, AWS SDK, IBM Watson, Azure SDK, Rev.ai SDK
- Utilized various online transcription engines to accurately transcribe children's voices, enhancing understanding through speaker diarization. Extracted detailed linguistic insights from TalkBank's HomeBank dataset, encompassing transcript conversion, frequency and sentiment analysis, statement length trends, and topic identification for comprehensive analysis.

Chicago Public Libraries Data Analysis | GitHub | July 2023

- Tools & Technologies: Python, Pandas, NumPy, Seaborn, Matplotlib, Folium, Scipy.stats, Geopandas, Shapely
- Performed spatial analysis with Folium to visualize the distribution and utilization of Chicago libraries, uncovering a decline in visitation post-2020. Validated this observation through rigorous statistical methods including hypothesis testing, outlier detection, and appropriate treatment measures.

IMDb Data Analysis | GitHub | June 2023

- Tools & Technologies: Python, Pandas, NumPy, Seaborn, Matplotlib, Scikit-Learn
- Conducted comprehensive analysis on a vast IMDb dataset covering 1940 to 2020, revealing trends in title releases, viewer preferences, and runtime dynamics. Categorized titles by type and genre, unveiling prevalent title types and viewer genre preferences over time. Utilized regression models (linear, polynomial, and random forest) to identify factors influencing viewer ratings accurately.

Business Inquiry to Research Solution | GitHub | April 2023

- Tools & Technologies: R, Dplyr, Ggplot2, Tidyr, Corrplot, Stargazer
- Analyzed small business data to discern federal aid uptake patterns, identifying key predictors like business size, location, and ownership status using logistic regression and decision trees. Proposed strategic approaches to enhance application rates and laid the groundwork for future randomized control trials, including power analysis.

Mining Software Repositories | GitHub | April 2023

- Tools & Technologies: Python, Pandas, NumPy, Seaborn, Matplotlib, PyDriller, GitHub API
- Utilized PyDriller and the GitHub API to systematically extract data from GitHub repositories, including commits, issues, and code structures. Leveraging visualization and metric computation techniques, offered detailed insights into repository evolution, providing actionable intelligence for software development stakeholders.

EasyID Data Identification Platform | GitHub | Dec 2022

- Tools & Technologies: C#, .NET, Blazor, Regular Expressions
- Developed EasyID, an interactive web application based on .NET technology, enabling data analysis and identification. EasyID features a user-friendly interface allowing data submission in various formats with real-time feedback. Leveraging the Blazor framework, EasyID combines .NET's backend prowess with modern web interactivity for efficient data processing. Implemented using object-oriented principles in C#, ensuring scalability and maintainability of the core logic and data processing components.

Spatial Analytics Exercise | GitHub | March 2022

- Tools & Technologies: R, Sp, Sf, Rgdal, Rgeos, Tidyverse, Tmap, Leaflet, GeoDa
- Consolidated diverse datasets from Chicago, using R for healthcare and socioeconomic analysis. Transformed data into shapefiles for advanced spatial visualization and utilized GeoDa for clustering and modeling. Geo-visualized city metrics to reveal spatial patterns linking health outcomes with socioeconomic conditions. Insights offer valuable guidance for community health initiatives and policymaking efforts.