

ANDRÉS CAMILO CARVAJAL S.

1105 Shiloh Ln Kennesaw GA, USA.

levasoftwaredata@gmail.com | (+1) (470) 926-3159 | [linkedin.com/in/andresca94/](https://www.linkedin.com/in/andresca94/) | andresca94.github.io

EXPERIENCE

SharpestMinds

Remote, USA

Data Science Fellowship with Mailchimp's Manager Data Scientist

Sep. 2022 – Present

- End to end Data Science, covering all the cycle of the data, state of the art models for machine learning and deep learning.
- Went over Data Science fundamentals such as creating models using supervised and unsupervised datasets.
- Researched state-of-the-art NLP technologies such as BERT and got a good understanding of the Transformer architecture.
- Implemented model explainability techniques such as SHAP.

CAREL USA

Roswell, GA

Software Developer, Automation Engineer

Jun. 2022 – Present

- Designing and developing solutions for HVAC control systems.
- Responsible for the design and development of HVAC/R software solutions using programmable controllers and established product platforms.
- Support and troubleshooting to customers for CAREL products and software solutions.
- Deployment and production full stack applications with Programmable Controls and Electronics.
- Time Series Forecasting using PROPHET and SARIMA for revenue data.
- HTML, CSS, JavaScript, Python, and structured text.

MR INGENIERIA

Bucaramanga, Colombia

Consultant

Dec. 2019 – July 2020

- Consultant activities of roads and oil facilities required in the development of the projects, for the Transportation Development Management of the Vice-Presidency of ECOPETROL S.A using GIS software.

Urban Development Institute of Bogotá D.C. IDU

Bogotá D.C, Colombia

Civil Engineer

Mar. 2019 – Nov. 2019

- Performed creation, conversion, translation, migration, and import of GIS data from various sources to the entity GIS (SIG-IDU).
- CAD and GIS to update existing records and make changes as necessary.
- Python scripting for data visualization, creation of reports, statistics, spatial data manipulation, and writing new files (txt, CSV).

DATA SCIENCE PROJECTS (more details at <https://andresca94.github.io/>)

Summarization and translation from YouTube videos using mT5 transformers.

- Multilanguage speech recognition using Whisper and Language identification with language detection library.
- Fine-tuned mT5 model encoder/decoder architecture to summarize and translate the recognized text from the video using xsum dataset.
- Rouge score of 0.25 for summarization.
- Deploy using Flask, Ngrok and Google Colab to run the virtual environment from Google on a server to use the power of GPU.

MasterCard stock price time series forecasting using LSTM and GRU

- Preprocessed dataset from May-25-2006 to Oct-11-2021.
- Built machine learning models to predict the stock price using LSTM and GRU. Pandas and Numpy for data manipulation, Matplotlib and pyplot for data visualization, scikit-learn for scaling and evaluation, and TensorFlow for modeling.

- GRU model got 5.95 rmse on the test dataset, which is an improvement from the LSTM model with rmse of 6.47.

Simple-Movie-Recommender and Content Based-Recommender from IMDB movies dataset

- This dataset contains IMDB Top 250 lists from 1996 to 2020 with every movie's basic information; release year, ranking, score, stars, etc.
- This project cover 2 approaches, the first simple recommender does the following: Decide on the metric or score to rate movies on. Calculate the score for every movie. Sort the movies based on the score and output the top results. The second approach build a system that recommends movies that are similar to a particular movie to finally create the recommendation based on the following metadata: the 3 top actors, the director, related genres, and the movie plot keywords.
- Compute pairwise cosine similarity scores for all movies based on their plot descriptions and recommend movies based on that similarity score threshold.
- Tokenized, vectorized text data using TF-IDF and cosine similarity to get a content based recommender and credits, genres, and keywords based recommender.

Multiclass prediction and clustering for music genre

- Data cleaning, handling entropy, feature engineering and data visualization.
- Encoded categorical data and preprocessed dataset. Trained XGboots, Random Forest and Logistic Regression, checking accuracy and confusion matrix to select the best model.
- Feature importance, ROC curve and SHAP values for explainability. Found the appropriate K-values with elbow method, balanced the dataset after feature selection and evaluated K-means visually with PCA.

NLP-Flask Language Prediction App

- Converted categorical variables to numerical, iterating through all the text and using regular expression to clean the data, train-test splitting.
- Model creation and prediction using Multinomial Naïve Bayes, confusion matrix, vectorizer and inverse transform to get the language prediction.
- Deployed using Flask, HTML and CSS.

Motor-Colission-App-Streamlit

- Loaded and cleaned the Motor Collision in New York City dataset. Creation of a 3D map and data visualization to respond the questions “Where are the most people injured in NYC?” and “How many collisions occur during a given time of day”.
- Breakdown by minute and affected type Pedestrians, Cyclist, Motorist. Front end using Streamlit.

CERTIFICATIONS

- **University of Toronto** – Coursera - 5 Courses - GIS, Mapping and Spatial Analyst tools Specialization Certificate.
coursera.org/verify/specialization/AMXMKL9G5FT6
- **University of Michigan** – Coursera - 5 Courses - Python for Everybody Specialization Certificate.
coursera.org/verify/specialization/MR4M59AABVW4
- **University of Michigan** – Coursera - 5 Courses - Python 3 Programming Specialization Certificate.
coursera.org/account/accomplishments/specialization/OFK6SAO89PK6

EDUCATION

ANDES University UNIANDES

Bogotá D.C, Colombia

Master of Science in Civil Engineering (STEM), Emphasis in Geotechnical Engineering

Jun 2021

Advisor: PhD. Nicolas Estrada Mejia;

Thesis: “Why the landslide size is a characteristic of fractal behavior?” <http://hdl.handle.net/1992/52990>

Coursework: Statistics, slope design workshop and containment structures (Data visualization), workshop of numerical tools in geotechnics (FORTRAN), continuum mechanics and models constitutive (MATLAB and Python), alternative energy, modeling and behavior of pavements (MATLAB), granular media physics and mechanics, physical geography (GIS).

Pontifical University UPB

Bucaramanga, Colombia

Advisor: PhD. Sandra Rocío Villamizar;

2017

Bachelor of Civil Engineering;

Thesis: “Waste utilization scheme common solid points for the UPB under the concept of circular economy” 5.0/5.0

SKILLS

Programming: Python (NumPy, Pandas, Scikit-learn, TensorFlow, PyTorch), SQL, MATLAB, GIT, BASH.

Front-end: HTML, CSS, JavaScript, Vue JS, Streamlit.

Back-end: Flask.

Visualization and Statistical Software: Excel, Python (Matplotlib, Seaborn), Figma, Adobe PS, ArcGIS.

Machine Learning: Regressions, Random Forest, XGBoost, Unsupervised Learning (Clustering, PCA), Deep Learning