Angel Santana Hernandez
Data Science | Data Engineer | Machine Learning

✉ ansa2343@colorado.edu
in @Angel_Santana_Hernandez
⬡ @Angel_Santana_
☎ 7202273047

SKILLS

Languages
- Python
- PySpark
- SparkSQL
- MySQL
- R

Visualization
- Tableau
- Streamlit
- Python Libraries
- R Libraries

Environments
- Jupyter Notebooks
- Git/Github
- Google Colab
- Visual Studio

Cloud
- Docker
- AWS EMR
- AWS S3
- GCP
- Hadoop

Competencies
- Machine Learning
- Data Preprocessing
- Data Mining
- Research/Analysis
- Agile methodology

EDUCATION

M.S. BUSINESS ANALYTICS
University of Colorado - Boulder
July '21 - Expected Graduation May '22

B.S. COMPUTER SCIENCE
University of Colorado - Boulder
Aug'16 - May '21

EXPERIENCE

Data Engineer-Machine Learning/ Team Lead- **Class Project** | Mar. 22 - Apr. '22
- Led a team of 5 through Preliminary Data Analysis, Exploratory Data Analysis, and Prediction Modeling. For our data, we worked with Used Car Sale Prices that had over 450 thousand rows of data and 16 columns. Our goal was to create machine learning algorithms that would predict car sale prices.
- **Responsibilities**: I was responsible for data cleaning, data transformation, and machine learning implementation.
- Using data cleaning/data transformation techniques, created an ETL pipeline that will allow the team to have the right data for exploratory data analysis. This included creating new columns, dropping NA values, and modeling columns.
- After EDA, using PySpark's machine learning library, I created a function that would do data preprocessing (indexing, encoding, assembling, and pipelining). Upon this, 2 different machine learning algorithms were created (Linear Regression and Gradient Boosting Tree Regressor) that would be used for predictions.
- See results here.
- Technologies/Frameworks used:
  - Jupyter-notebook **|** Git/Github **|** Python/PySpark **|** AWS EMR/S3 **|** Docker

Data Scientist Intern - **Fortune 500 Company** | Jan. '22 - Feb. '22 | (under NDA)
- In a team of 5, we were tasked to analyze HR data to use for model prediction on employee churn. Our goal was to see which features indicated that an employee would quit. The data that given consisted of over 89 thousand rows and 20+ columns.
- After data cleaning and data transformation as a team, we came to a conclusion on a new dataset. I then worked on exploratory data analysis to get a preview of our new data and common trends.
- Once there was enough analysis, I implemented a **Gradient Boosting Classifier** machine learning algorithm.
- We then compared our models and chose the best classifier based on Precision, Recall, and Accuracy.
- Technologies/Frameworks used:
  - Google Colab **|** Python

Machine Learning Research/Team Lead - **Class Project** | Oct. 2021 - Dec. 2021
- Led a team of 3 through Exploratory Data Analysis and research methods using Machine Learning Analysis. We worked with the Boston Housing Dataset to explore which features are highly correlated that could contribute to a high prediction accuracy in housing value.
- Through our EDA phase, we looked at statistical analysis of our features provided, correlations of features against our predictor variable, and looked at any outliers that might affect our accuracy on our predictive models.
- After finding our correlations and features to include in our machine learning analysis, I then implemented 3 different linear regression models after going through tree analysis, where we found which features contributed more to our target variable.
- The end goal of this project was to implement machine learning algorithms with different parameters and notice how the coefficients of our features would change, and improve on that.
- You can view the teams final results here.
- Technologies/Frameworks used:
  - Jupyter notebooks **|** Git/Github **|** R

SQL Data Analyst/ Team Lead - **Class Project** | Sept. '21 - Dec. '21
- Led a team of 5 through Data tranformation and Exploratory Data Analysis. Our goal was to provide a property investor with analysis on Airbnb data in the metro area. Our goal was to determine use different queries to find analysis that can contribute aid to this investor.
- During our data transformation phase, we created new data files that would allow for better analysis. We got rid of features and instances that didn't add any value.
- After creating our new data, using MySQL, we created queries to get insights on the property in the metro are. You can view the final compiled results here.
- Technologies/Frameworks used:
  - Visual Studio (MySQL) **|** MySQL Workbench **|** MySQL (Data Querying) **|** Git/Github **|** Python (Data transformation)