|  |  |
| --- | --- |
| DAVID **CASTILLO** | Brooklyn, NY  718.415.6767  dcast822@outlook.com |

|  |  |  |
| --- | --- | --- |
| **LinkedIn:** [linkedin.com/in/davidcastillo822](https://www.linkedin.com/in/davidcastillo822/) | **GitHub:** [github.com/dcast822](https://github.com/dcast822) | **Portfolio:** [datadavidc.com](https://www.datadavidc.com/) |

**>LANGUAGES & TECHNOLOGIES**

* **Proficient:** Python, Pandas, NumPy, SciPy, Seaborn, Matplotlib, Machine Learning, Scikit-Learn, NLP, NLTK, Statsmodels, Deep Learning, Keras, Tensorflow, SQL, Data Structure & Algorithms, dbt, BigQuery, Looker Studios, Redash, Google Cloud Products, Redshift
* **Exposure:** Apache Spark, Streamlit, Flask, Heroku, Amazon Web Services, Tableau

**>BUSINESS & STATISTICAL CONCEPTS**

* A/B Testing, Experimental Design, Hypothesis Testing, Statistical Analysis, Business Intelligence, Visualization

**>PROFESSIONAL EXPERIENCE**

**Dow Jones** **| Senior Data Analyst | New York, NY** July 2023 – Present

* Automated various repeated procedures into modular tables, utilizing BigQuery, Scheduled Queries and SQL, resulting in modular tables that helped increase query speeds by ~3X, also reducing SQL code lines written.
* Developed reporting dashboard for all programmatic advertising data, utilizing GAM API, Python, Airflow, BigQuery, SQL, Looker Studios, which utilizes data pipeline to help track various disparate revenues cut across various segments from all programmatic channels. Dashboard utilized by various stakeholders including executives, sales, planning and operational teams to align source of truth across these departments and personnel.
* Helped reconcile data discrepancies, utilizing GAM UI, GAM API, Python, BigQuery, SQL, Looker Studios for Tax Revenue and Video Advertising Dashboards, reducing discrepancies from ~20% to ~.05% due to fixing how the data was aggregated and then joined.

**Kaiyo** **| Analytics Engineering Intern | New York, NY** January 2023 – July 2023

* Provided ad hoc analysis & reporting, using DBT, SQL (Redshift) and Redash, to aid operations team in optimizing durations through various activities segments related to receiving and maintaining incoming furniture to make ready for sale, leading to a decrease of overall durations by ~12%.
* Built, documented & tested data models, used for various departmental analysis such as addressing overpayments from revenue share from furniture sales, utilizing DBT, Stitch, and SQL, resulting in maintaining 0 overpayments to clients.
* Constructed queries via large datasets from warehouse, for alerts, dashboards, and data visualizations, utilizing DBT, SQL(Redshift), and Redash, to help departments improve processes such as modeling routes and their on-time delivery frequency, helping to increase on time delivery performance from ~92% to ~97% on time rate.

**T-Mobile** **| Manager | New York, NY** July 2013 - October 2021

* Conducted workshops on best practices for sales, using Dale Carnegie and Kaizen principles, leading to ~15% less escalations from the prior year and a decrease in ~$1.5K in credits.
* Established key performance metrics for the sales team, using sales trends through Power BI data, resulting in 5 goals met at ~105% of target.
* Ensured compliance with the Operations Department leading to the resolution of discrepancies that shifted losses from my company to the logistic company in the average amount of ~$15K a year.
* Conducted training, using Dale Carnegie principles, and Kaizen, leading to 40+ personnel having promotions or improving ability to meet their targets.

**>PROJECT WORK**

**Housing Sales Price Prediction with Machine Learning | Data Scientist |** [**Code URL**](https://github.com/dcast822/Housing_salesprice_prediction_with_Machine_Learning)2022

*Built Linear Regression model that predicts housing prices using ~30 different features.*

* Achieved an R2 score of 90%, through processing outliers & missing data, using Python, Scikit-Learn, Pandas & Numpy.
* Used regression analysis to gain insights such as with all else held constant, a house in the Greenhills neighborhood increases value by $102K, using Python, Scikit-Learn, Pandas, and NumPy.

**Credit Risk Analysis Using Classification Methods | Data Scientist |** [**Code URL**](https://github.com/dcast822/credit_risk_analysis)2022

*Iterated through classification models that predict whether a borrower will default or not.*

* Achieved AUC Score of 95% & Recall of 72%, indicating how well you can predict from within the population who will default, using Python, Scikit-Learn, Pandas and Numpy.
* Used XGBoost model’s information gain to get access to important features such as; loan grade, home ownership status, and Loan as a percent of income, using Python, Scikit-Learn, XGboost, and Matplotlib.

**>EDUCATION**

**Data Science Bootcamp,** *General Assembly* Feb 2022

**MBA (Accounting),** *University of Arizona Global Campus* March 2021

**Bachelors (Political Science),** *CUNY Lehman*  May 2010