Daniel A. Legorreta

https://dlegor.github.io

Email: d.legorreta.anguiano@gmail.com
Mobile: 512-871-9006

Work Experience

Data Engineer/ Machine Learning Engineer - *Automated FruitScouting, Inc*

09/2021 - Present, Austin,Tx.

- Developed and improved various segmentation and object detection models, especially in the problematic case of small and confusing segmentations, as well as in the detection of numerous very small objects in the same image. The improvements to the models were reflected in performance, which increased from 0.73 to 0.92 in segmentation. In object detection, I detected data leakages and achieved an accuracy improvement of up to 20%. These models formed the basis of the company's products and services.
- Engineered and implemented a comprehensive data version control system and an end-to-end MLOps pipeline, streamlining dataset consolidation, model training, and error analysis, resulting in a continuous model improvement cycle
- Created multiple scripts and codes to solve various problems, such as the organization of data via API, the processing and organization of images, the analysis of inconsistencies in data with labels, implementation of unsupervised models to detect irregularities in the images or anomalies, implementation of OCR to detect anomalies in certain sets of images, implementation of different computer vision techniques to evaluate the complexity of the images, developed models to estimate relationships in different business cases (regressions and classifications).
- Maintenance and development of pipelines and deployment of models in production and experimental environments.
- Developed a small system for performing semi-annotations. I used the models in production to help create new annotations and thus reduce the time it takes to create new datasets, thus ensuring a continuous model improvement cycle.

Independent Consultant - *Machine Learning Consultant*

07/2017 - 09/2021, Austin,Tx & Latam.

- Developed for T7Group, which is a partner of AT&T in LATAM, a system to process multiple time series data and built forecasting models for the monitoring and maintenance of the AT&T network, the system helped them reduce the detection that alarms (outliers) and allowed have perspective of the root cause of the possible problem in the future and improve the attention of the resources that the company allocates for the maintenance of its network, which represented a saving for AT&T.
- Developed for T7Group and Airbus, a system to process multiple time series data in real-time and I built several models for anomaly detection. That system was part of the various components that these companies created for the Mexican Army. The system was deployed throughout Mexico and helped the Army Network improve the monitoring and detection of strange situations in the network's operation.
- Developed packages and systems for UNDP Accelerator Labs(United Nations) to process the texts of all the projects in the Mexican Government, these projects are where the government spends for the implementation of public policies. The system helped detect how government officials report progress, what type of errors they make when reporting, how they had biases in their reports, and it made it easier to detect what were the possible causes of non-compliance with the goals in government programs. This had such an impact that in 2021 the government changed the mechanism through which it guides officials, as a result of the findings that the system was allowed to obtain.
- Developed an application for the company Ixulabs, which allows identifying from a code of letters and numbers generated by one insurer to another related code for 12 other insurers, the system had 3 layers with different types of models. The first was a set of translators (seq2seq models), the second a combination of embeddings and knn, and the third an adaptation of the BM25 algorithm to find the most similar ones. This application allowed the company to detect its most similar codes in the 24 thousand codes. This functionality is granted to clients of an insurance agent platform, as a new service.
- Implemented NLP techniques and algorithms for customer segmentation, topic detection, classification, semantic textual similarity, sentence pair modeling and developed chatbots with different components (or actions) to solve data quality and customer service problems for various cases of business.
- Developed a system to analyze pdf files through OCR, NLP, and Deep Learning algorithms to extract text, tables, tags, and specific keywords. This to feed other systems for searches and improve the handling of digitized information.
- Implement and develop a set of pipelines and models for a Latin American bank, which has 30 million clients. These components and models had the objective of detecting the

probability of default on loans, helped the bank to improve its analysis of credit risk and the management of its clients.
- Created and deployed various models for classification and regression problems for tabular data from various business cases.

Santander Bank - *Data Scientist & Big Data Consultant*

10/2015 - 07/2017, Mexico

- Developed a system to analyze and estimate the behavior of the Bank's clients who manage certain types of accounts. These accounts are used by large companies, which represented a business option for the bank, to better manage that capital in the international financial markets while the companies accepted. The system helped better decide which companies to offer the offer to and which to follow, such as possible accounts that could be withdrawn from the bank.
- Created code to process the logs of the bank's mobile application in order to detect intrusions in the application and I implemented an anomaly detection layer. This helped detect failures and intrusion attempts in the application, allowing the security team to improve the patches and have the routes or ips from which the application was attempted to be hacked identified.
- Analyzed the information and estimated the probability of default and expected loss from the credit bureau and bank data for clients handling business accounts.
- Designed dashboards in Tableau-Hadoop for project management reports for mobile apps and other business cases.

CES Consultant & Battersea Consultant — *Senior IT Consultant*

08/2015 - 06/2016, Mexico

- Created the codes to estimate poverty in Mexico 2016 in R, SPSS and Stata, it was a project developed for CONEVAL under CES consultancy.
- Developed the codes to monitor the communication between the applications and the server that allowed analyzing the activities among the users.

Softtek — *IT Consultant & Statistical Consultant*

01/2012 - 06/2014, Mexico

- Built predictive models to control events with projections by day, week and month.
- Created a decision tree to facilitate the allocation of events to the service desk.
- Automated the daily event reports, which reduced the time from 1.5 hours to 10 minutes.

Education

- Bachelor of science in Physics and Mathematics, ESFM-IPN Mexico.
- Master level courses in Mathematics, Cinvestav Zacatenco-IPN México.
- Research Assistant in Mathematics and Complex Systems Cinvestav and UPIITA Mexico.

Skills:

Machine Learning, Deep Learning and NLP:Tensorflow,Pytorch,Keras, Numba, Numpy, Scipy, Dask, Spacy,SageMaker,PySpark,Pandas, Polars,transformers,Datasets,Scikit-Learn,XGBoost,optuna.
Big Data:Spark, Scala, Solr, Druid, Beam, Kafka,Airflow,BigQuery,HBase.
Other: R,Python,Linux, AWK, Cloud Computing (GCP and AWS), Docker, Kubernetes, Serverless Computing,Git, Apache Superset,Cvat,VertexAi,Kuberflow.
Certification:

- AWS Certified Machine Learning - Credential ID:VSNEVC02JFBEQEG3
- Advanced Machine Learning with TensorFlow on Google Cloud Platform Specialization

Relevant Projects (more at https://dlegor.github.io/projects.html):

Google Analytics Customer Revenue Prediction (Kaggle data science competition) Ranked among the **6% top worldwide**.
Home Credit Default Risk(Kaggle data science competition) I finished in the **top 12% worldwide**, of the most popular competition in Kaggle, with more than 7000 participants.
Power Laws: Forecasting Energy Consumption (DriveData data science competition) Ranked in the **top 4% worldwide**,the problem was to analyze and forecast many time series with different features and time windows.
Research Projects:
MicNet toolbox: Visualizing and unraveling a microbial network
Minería de texto en el sistema de evaluación del desempeño
Patent - Systems and methods for automated Crop Load Management