Data Science For Social Good
Summer Fellowship

Modeling Prices in Public Contracting

Beck DeYoung *Fellow*
Maria-Cristiana Gîrjău *Fellow*
Dushant Gohri *Fellow*
Ethan Lee *Fellow*
Robert Hager *Technical Mentor*
Daniel Townsend *Project Manager*

DNCP DIRECCIÓN NACIONAL DE CONTRATACIONES PÚBLICAS

Background

Partner National Directorate for Public Contracting in Paraguay (DNCP)

The DNCP **oversees public spending**, **maintains a registry** of transactions, and tries to **identify corruption** or other suspicious activity.

Project Goals

Being able to estimate reasonable prices of goods and services helps the DNCP **detect overspending and corruption**

TWOFOLD AIM:

> Train models to **predict item prices**
> Establish a **reusable data science pipeline** for the DNCP's future use

Data at a Glance

Historical data from **2010 to 2021**

Approx. **3.3 million** goods and services

40 predictors chosen from **>180 variables**

Data Pipeline

Key Features of the Pipeline

> **Flexibility:** user-friendly and fully customizable
> **Modularity:** self-contained at every step
> **Robustness:** handles edge cases and erroneous input

PIPELINE HIGHLIGHTS

Isolation forest to refine model focus by **removing extreme prices** (e.g., bridges, highways)

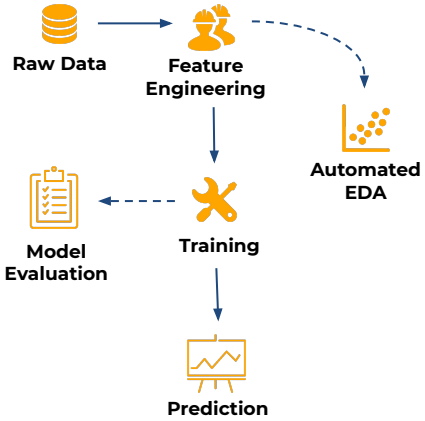
Convert currencies to PYG and **adjust prices for inflation** using the consumer price index

Contextual Indicators

ISSUE: Not enough information on the context/purpose of a purchase (e.g., scissors for school vs. scissors for hospital)
SOLUTION: Create indicator variables for various common situations

> Text mining on the **buyer** and **tender description** fields
> Identify a total of **9 buyer indicators** and **20 description indicators**

Workflow



Model Types

BASELINE: ordinary least-squares multiple linear regression

Train 2 competing model types:
> **XGBoost**
> **Random Forest Regression**

Highlight: automated tuning of model hyperparameters

Model Training

Model segmentation: train/evaluate 2 separate models for goods and services
Findings from previous iterations: low prices have very high percentage error, so train models only on items above $2

Results

Best model: XGBoost

> Price variability accounted for by the model is **77% for goods** and **80% for services**
> Root median squared error of **$12 for goods** and **$33 for services**
> Median percentage errors of **48% for goods** and **53% for services**

Looking Forward

The data pipeline can be used by the DNCP to **train more models**

Further refinement to find the sweet spot between excluding **low prices** and **high prices**

Construct **confidence intervals**

This work was done during the Data Science for Social Good Summer Fellowship at the University of Warwick and Ludwig Maximilian University of Munich