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*

Background



National Directorate for Public Contracting in Paraguay (DNCP)

The DNCP oversees public spending, maintains a registry of transactions, and identifies 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:

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

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



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)



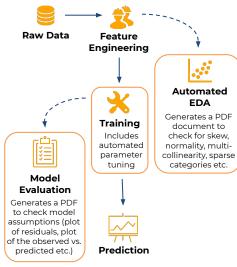
Convert currencies to PYG and **adjust prices for inflation** using the consumer price index

Contextual Indicators

ISSUE: Not enough information on the context/purpose of a purchase (e.g., scissors for school vs. scissors for hospital) **SOLUTION:** Create T/F 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 model types: > **XGBoost**
- > Random Forest Regression

Model Methodology

- > Train/evaluate 2 separate models for goods and services
- > Since models including low price data have very high percentage error, focus on items with mid-range prices

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 **most appropriate model focus** (e.g. low prices and high prices)



Confidence intervals or quantile regression