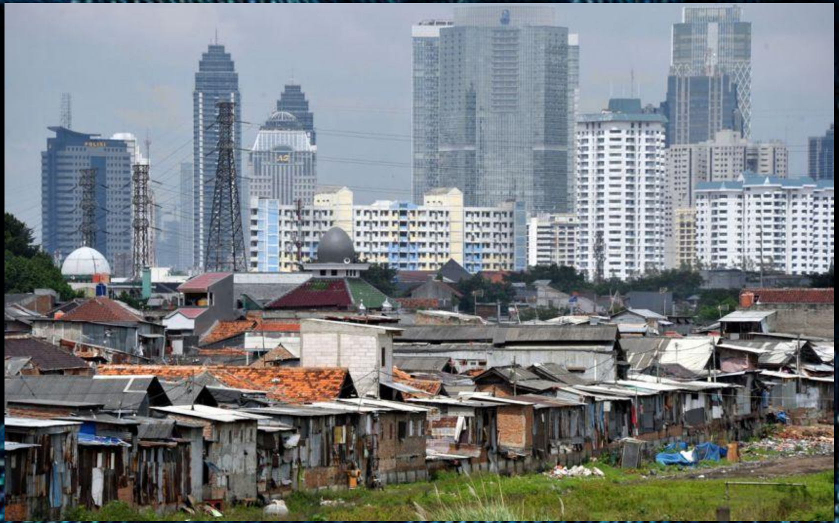


Informal settlement Identification

… Using Real Estate & Geo processed data

Kathirvel Kumararaja
Feb 25, 2020

Kathirvel Kumararaja

Vice President - Operations,
DevJee Inc.

Credentials:

- Data Scientist with 20 + years of IT consulting experience.
- Holds an Engineering degree and and MBA.
- Experience in handling data analysis for multi-billion dollar capital development projects
  - Burj Khalifa tower in Dubai
  - Pentagon Renovation project, Arlington Virginia.

Agenda

- Problem Statement
- Bird’s eye view
- Findings
  - Data overview - EDA
  - Model evaluation
- Conclusions and recommendations

Problem Statement

Business Objective

To deliver Effective economic and social aid, non-government organizations require detailed maps of the locations of informal settlements.

Challenges

Informal settlements are home to the most socially and economically vulnerable people on the planet.

Desired Outcome

- Use public data and Geo processing to engineer data
- Train the models to predict the informal settlements.
- Evaluate the model using Accuracy as the criteria

Birds eye view

Qualitative data 1

- Two Real Estate Data sets
    - ■ Brazil and Sao Paulo

Qualitative data 2

- Time scale
    - Nearly three years

Qualitative data 3

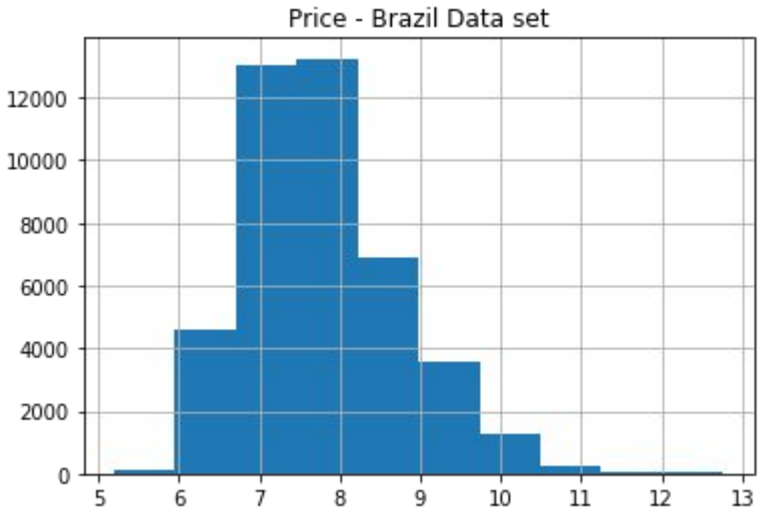
- Baseline Score
    - 0.78

Qualitative data 4

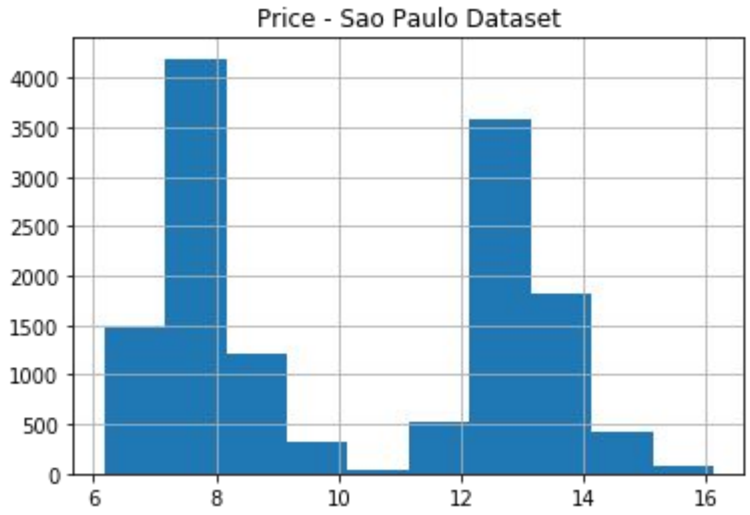
- Models / Classifiers Explored
    - Logistic Reg, Random Frst, Extra trees.

Data Overview - EDA

EDA - Price fields



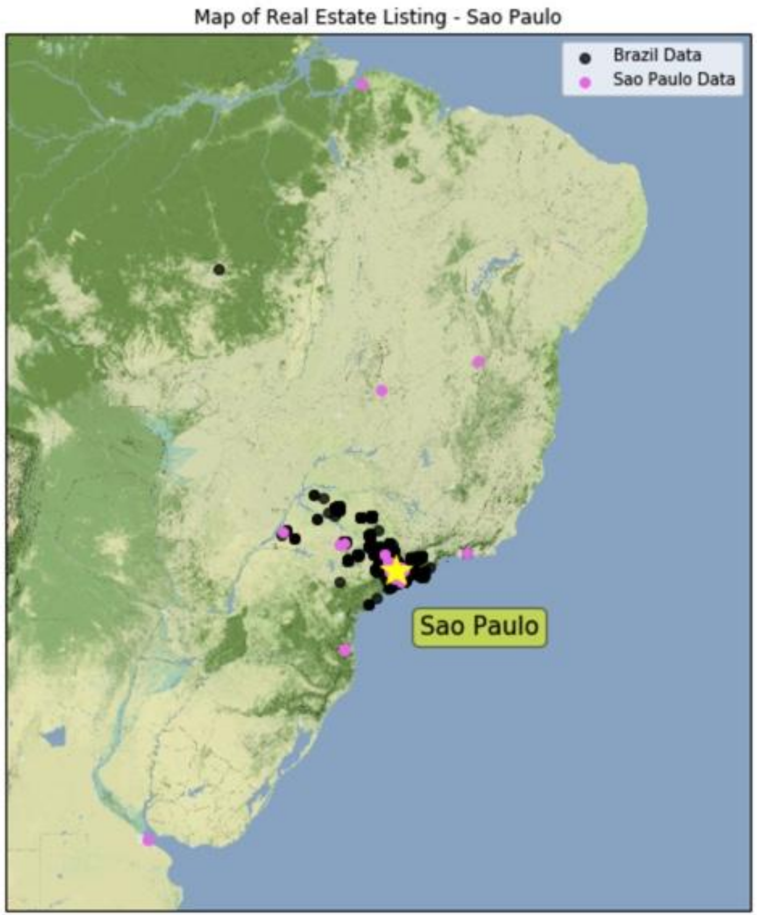
Price - Brazil Data set



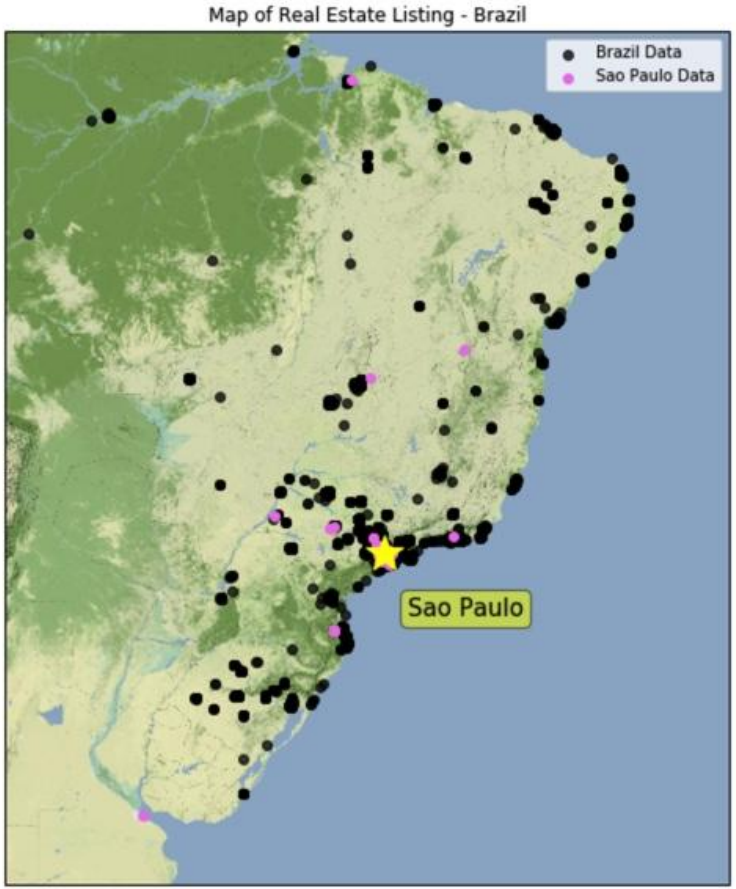
Price - Sao Paulo Dataset

EDA - Real Estate Listings - Using Carto py

T



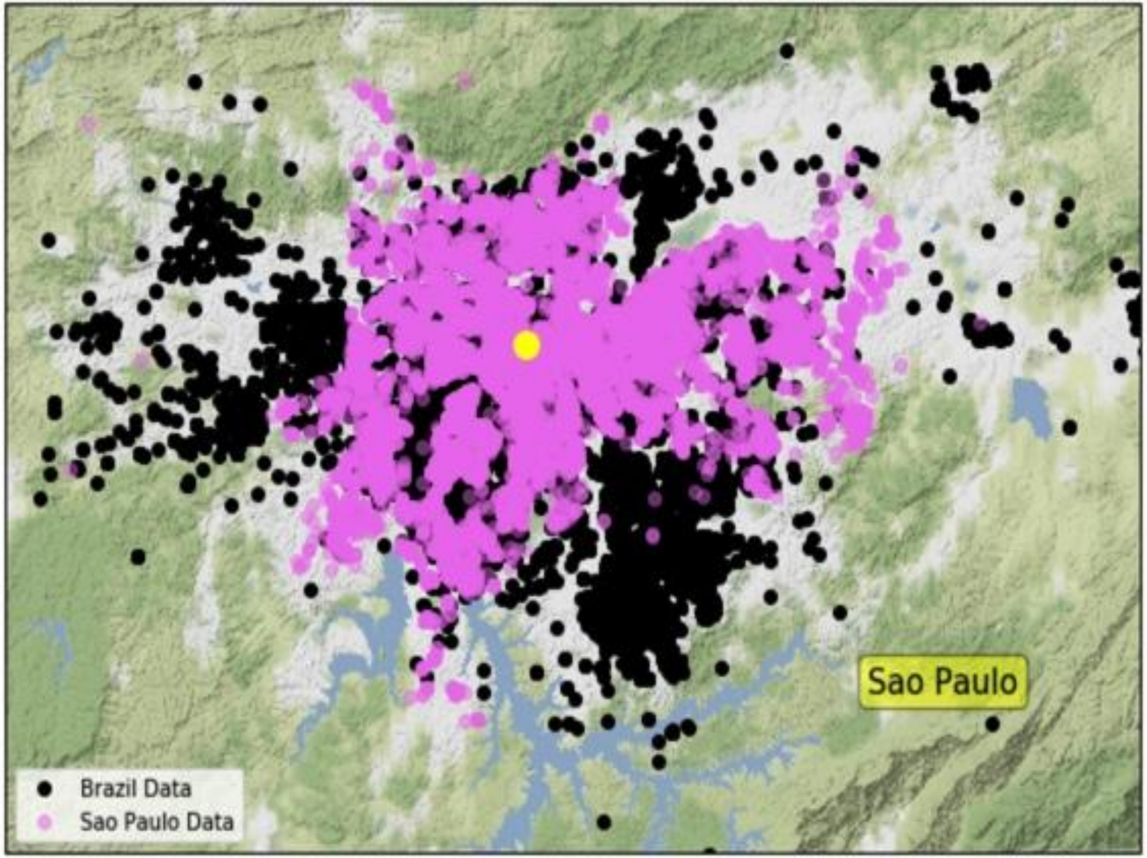
Map of Real Estate Listing - Sao Paulo



Map of Real Estate Listing - Brazil

EDA - Real Estate Listings - Sao Paulo

Combination of the two datasets allow us to have more data covering the whole city.



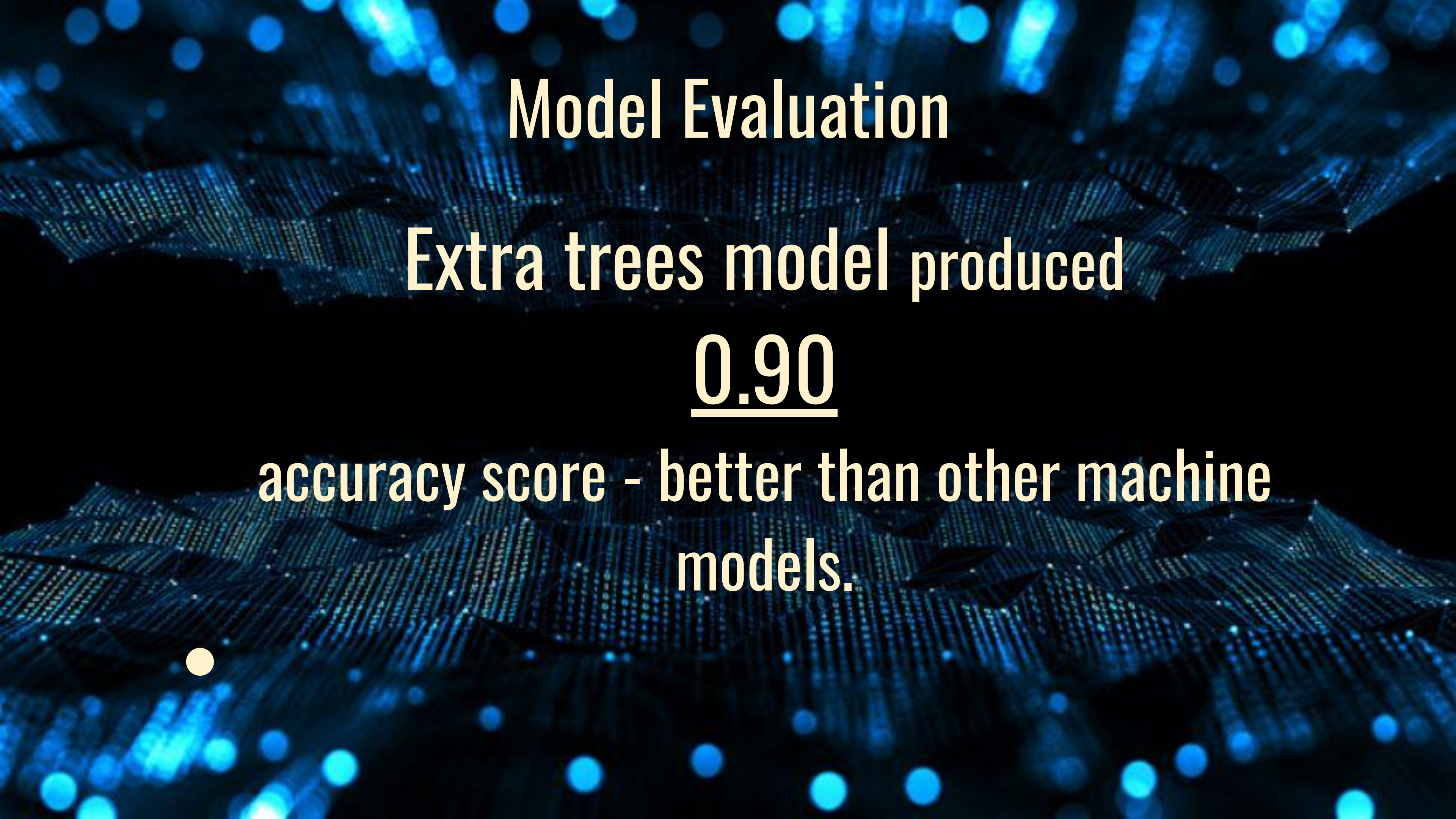
Model Evaluation

Criteria :

- Main Metric - Testing Accuracy
  - Confusion Matrix
  - ROC with AUC curve
  - Model coefficients

Model Performance - overview

Model	Train Score	Test Score
Base Line	0.78	
Logistic Regression	0.78	0.77
Decision Tree	1.0	0.90
Random Frst.	1.0	0.90
Extra trees	1.0	0.90
Voting clsfr.	0.99	0.87



Model Evaluation

Extra trees model produced

0.90

accuracy score - better than other machine models.

Model Evaluation - Confusion Matrix

- Accuracy = 0.90

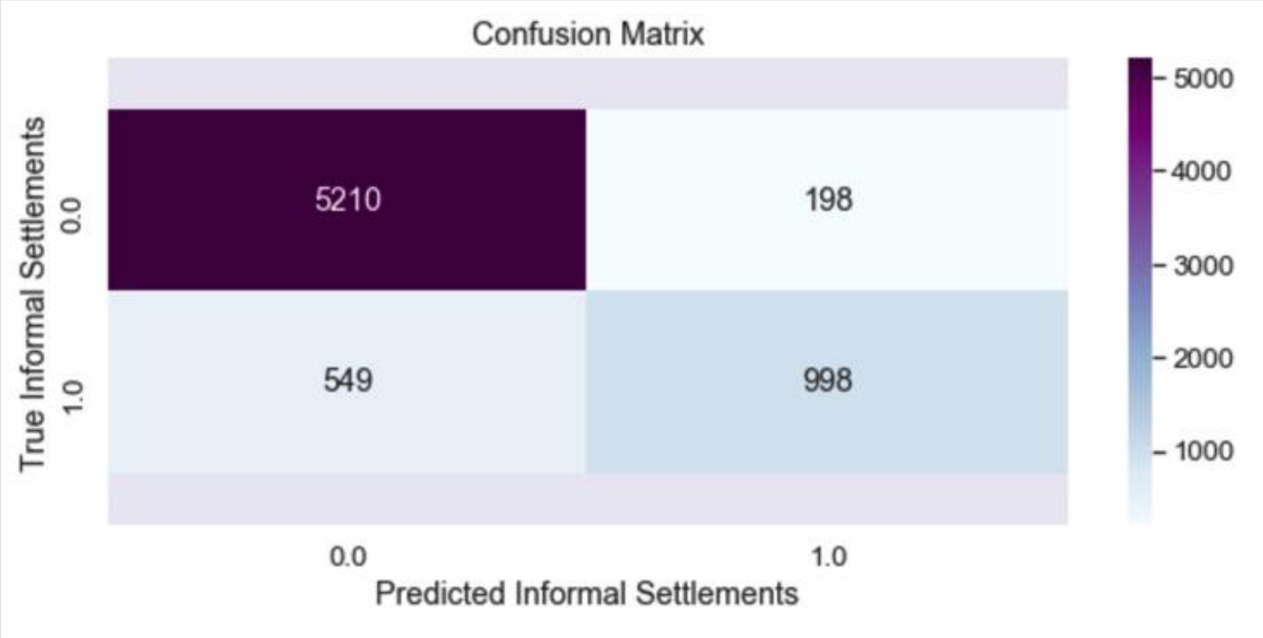
- Misclassification Rate = 0.12

- Specificity = 0.82

- Precision = 0.86

- Sensitivity = 0.93

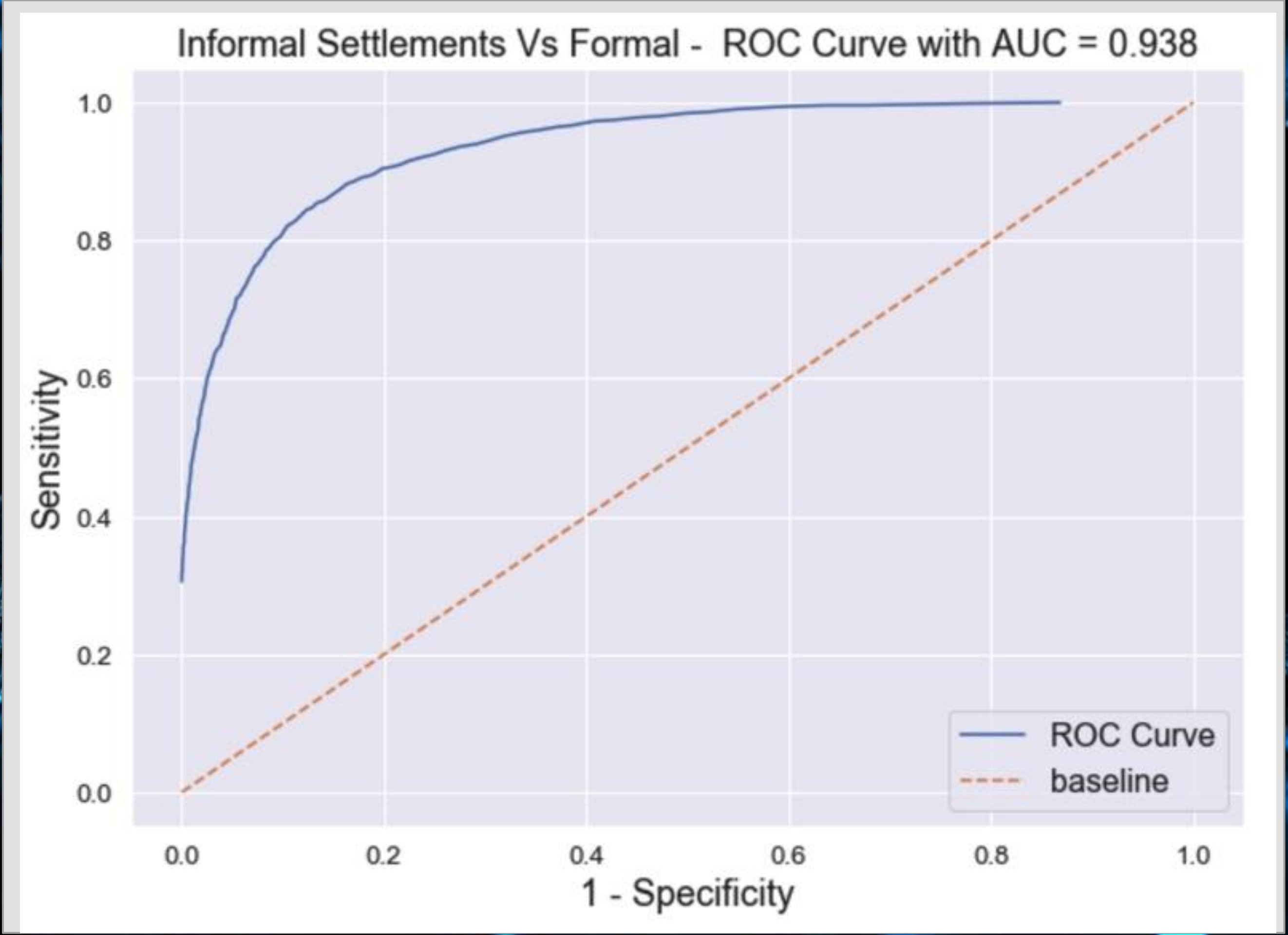
https://predictfavelas.github.io/kumar_predictions


Confusion Matrix

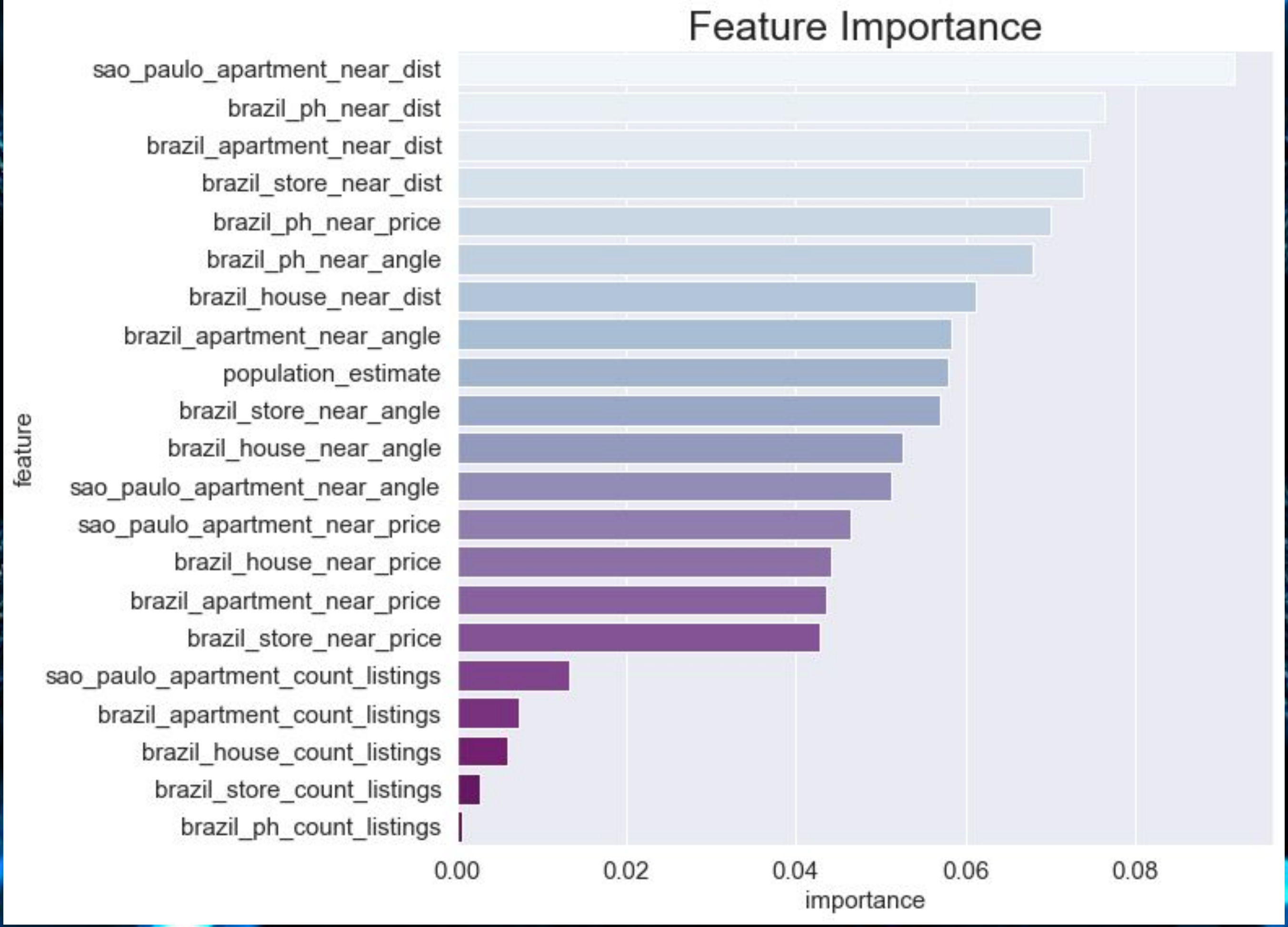
Model Evaluation - ROC AUC Curve

- ROC AUC of close to 1

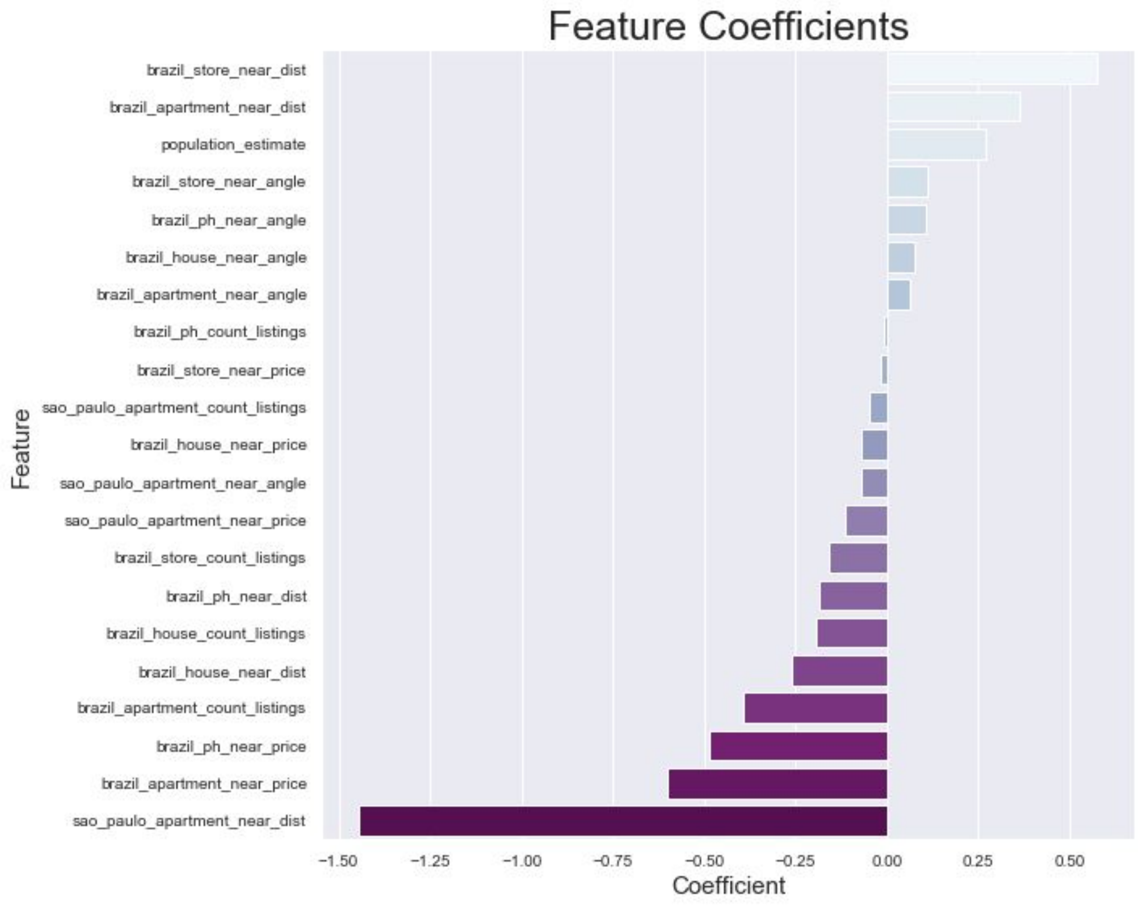
- Positive and Negative classes are perfectly separated



Model Evaluation - Feature Importance (Extra trees)



Model Evaluation - Model Coefficients



Conclusion

- Extra trees model performed the best (at 90% accuracy).
- Our model will help to differentiate the informal and formal settlements for aid agencies.
- Would like to get more data from other countries to train my model and bring down the variance.
- Would like to do better feature engineering using other real estate data and Geo-spatial data.

Thanks - Questions?

The Team

Kathirvel Kumararaja

Mahdi

Ambar

GA DSI 10 Cohort