City-Level Growth Potential Prediction

**Team:** Michelle Ma and Andy Zhao  
**Faculty Advisor:** Prof. Rama Ramakrishnan  
**CTI Advisor:** Ashok Mehta

### Project Context and Overview

**Problem Statement**

CTI Real Estate Research wishes to develop a Machine-Learning driven tool to identify European cities with top long-term future growth potential. This could help in facilitating the team's investment decisions. Utilize analytics to gain insights on factors impacting city growth potential, and determine:

- Cities with top growth potential across 8 European countries
- Most important indicators that impact future long-term potential

**Dataset Description**

- Economic drivers
- Knowledge economy
- Environment, social, & governance
- Country attractiveness
- Demographics
- Connectivity
- Liveability

**Project Objectives**

- Develop analytical approaches to predict long-term future growth potential for 600+ European cities:
  - Problem framing with Real Estate Research team
  - Select independent variables from existing dataset
  - Predictive Modeling
  - Enhance interpretability of results

- Assist Real Estate Research team in making investment decisions with rank of top growth potential cities

**March-April:** conduct initial exploratory analysis on the dataset of independent variables

**May:** Identify long-term growth potential indicators, establish baseline criteria, build first version of Linear Regression models

**June-July:** Build and evaluate first version of boosting models, select desired training & test time frames

**July-August:** Refine modeling approaches to include new independent variables, explore additional modeling approaches & extract insights

### Methods

**Direct growth potential prediction approach**

We want to predict future growth potential with data collected from the past and deliver our results in an interpretable manner. One method to do so would be using Linear Regression models that display the weight of each factor in making predictions. To leverage the extensive data sources and extract the most insights, we also used boosting models as an additional method.

\[
\text{Boosting model: Linear regression} \quad \text{Predict & compare: 1-year growth potential vs. 3-year growth potential}
\]

**Two-step implied growth potential prediction approach**

To gain an additional perspective, we experimented with a two-step modeling approach. In step one, we used boosting models to predict the level values of growth potential indicators. In step two, we derived the implied rate of increase from the corresponding level values.

**Selection of Performance Metrics**

To gain a comprehensive perspective on our models’ performance, we assessed the prediction outputs with a variety of performance metrics:

- **Accuracy in ranking of cities’ growth potential:** Mean Absolute Errors for top n ranked cities, Spearman's Rank Correlation
- **Comparison of predictions relatively to set criteria:** Binned Absolute Errors within 50 to 250 basis points, % predictions with opposite signs from true values
- **Comparison of performance against baseline model:** Training R² and Test R² values

**Evaluation of Modeling**

Comparison of performance between direct vs. two-step growth potential prediction models:

- Different performance evaluation metrics help us gain insights from different perspectives on our models. For instance, the Binned Absolute Errors plot shown below demonstrates that the direct model outperforms the two-step model through capturing more % absolute errors within 50 and 100 basis points.

**Top important factors impacting growth potential**

From boosting model and linear model outputs, we identified top important factors that impact city-level growth potential predictions and the common themes these factors belong to.

**Themes of top important factors:** The distribution of top important factors among the 7 themes of data helps CTI Real Estate Research team identify data sources that effectively predict cities with high growth potential

Additionally, SHAP value analysis enhances interpretability of model outputs by demonstrating how each factor positively/negatively impact prediction results.

**Recommendations for next steps**

- Explore with country-level prediction models for potential improvements in accuracy
- Continue the current fruitful journey of collecting data and leverage state-of-the-art boosting models. Further gains in predictive ability are likely to come from enhanced data rather than from more powerful models
- Include predictions from other independent sources into the dataset to elevate boosting model’s predictive power

**Our contribution**

- Built a comprehensive model pipeline, from feature engineering to evaluating output
- Provided diverse perspectives in prediction time frames and growth potential estimation techniques through a variety of modeling approaches and performance metrics analysis
- Enhanced interpretability of model outputs with analytical techniques such as SHAP analysis, communicate direct insights to all stakeholders
- Provide recommendations on future data enhancements, that can get seamlessly folded into the modeling process
- Handed out model that can be “cloned” and adapted to address other prediction targets of interests at CTI