Capstone Final Project Report:
Predicting Tech Startups in Los Angeles Metro Area

Wesley Hall
September 7, 2024

---

1. Introduction & Problem Statement

The objective of this project is to predict the future spatial distribution of technology startups in Los Angeles, based on business registration data, educational attainment metrics, and temporal factors.

By understanding how educational demographics and temporal trends influence business clustering, this analysis aims to inform targeted urban development and economic policies that foster tech growth in key areas of Los Angeles. As a growth industry, technology-related companies need to rely on an educated workforce to thrive , fully utilizing the city's world-class higher education infrastructure to power its future. Knowing the types of technology startups that will develop in the coming years will inform the allocation of energy and resources in coming years.

---

2. Data Overview

I originally intended to analyze all of Los Angeles County but, due to the large dataset (over 3 million data points), I filtered the data by zip code to focus on areas where tech companies tend to cluster, such as Santa Monica, Venice, Westside Los Angeles, and Downtown Los Angeles.

Data Sources:

1. **Business Registrations**: Dataset includes business registration information categorized by NAICS codes, zip codes, geographic coordinates (latitude, longitude), and registration dates.
    - **Source**: LA Active Businesses Dataset
2. **Educational Attainment**: Provides metrics on educational attainment (e.g., percentage of the population with bachelor's degrees or high school diplomas) by zip code.
    - **Source**: Census Educational Attainment Dataset

Timeframe:

Data was collected for the years 2011 to 2022. The incomplete data from 2023 and 2024 was excluded from the analysis.

Key Data Wrangling and EDA STeps:

- I stripped leading and trailing spaces from column names and ensured zip codes were formatted correctly as strings.
- The datasets were merged using zip codes. I ensured all necessary columns were converted to the appropriate types, and any missing values were handled.
- After filtering the data to focus on specific zip codes, I conducted exploratory data analysis (EDA), including generating summary statistics and plotting the distribution of tech startups over the years.

Key Preprocessing Steps:

- Filtered zip codes to focus on Santa Monica, Venice, and other key areas.
- Handled missing values and standardized the data.
- Applied Principal Component Analysis (PCA) to reduce feature dimensions.

---

3. Methodology

I employed a combination of supervised learning techniques and clustering methods to predict the locations and types of future tech startups in Los Angeles.

Modeling Approach:

- Dimensionality Reduction:
  - I used Principal Component Analysis (PCA) to reduce the number of features. The first principal component (PC1) explained 75.7% of the variance, allowing me to combine correlated features like educational attainment across age groups.
- Model Selection:
  - I experimented with Random Forest Classifier (RFC), Decision Tree Classifier (DTC), and Logistic Regression. I also combined models using ensemble methods (e.g., Stacking and Bagging), but these performed worse than the individual DTC and RFC.
  - Final Model: The Extra Trees Classifier outperformed all other models after tuning hyperparameters and using PCA, achieving an 80% accuracy.

Hyperparameter Tuning:

- I used GridSearchCV to fine-tune hyperparameters such as max_depth, min_samples_leaf, min_samples_leaf, min_samples_split and n_estimators.

- After testing ensemble methods, the best results were achieved using the Extra Trees Classifier with a tuned n_estimators=200 and no maximum depth (max_depth=None).

- Clustering: I applied **K-Means Clustering** to predict future business clustering based on geographic location (latitude and longitude) and business characteristics (NAICS codes, revenue, or size). This clustering analysis helped visualize where tech startups are likely to emerge in the future.

---

4. Results

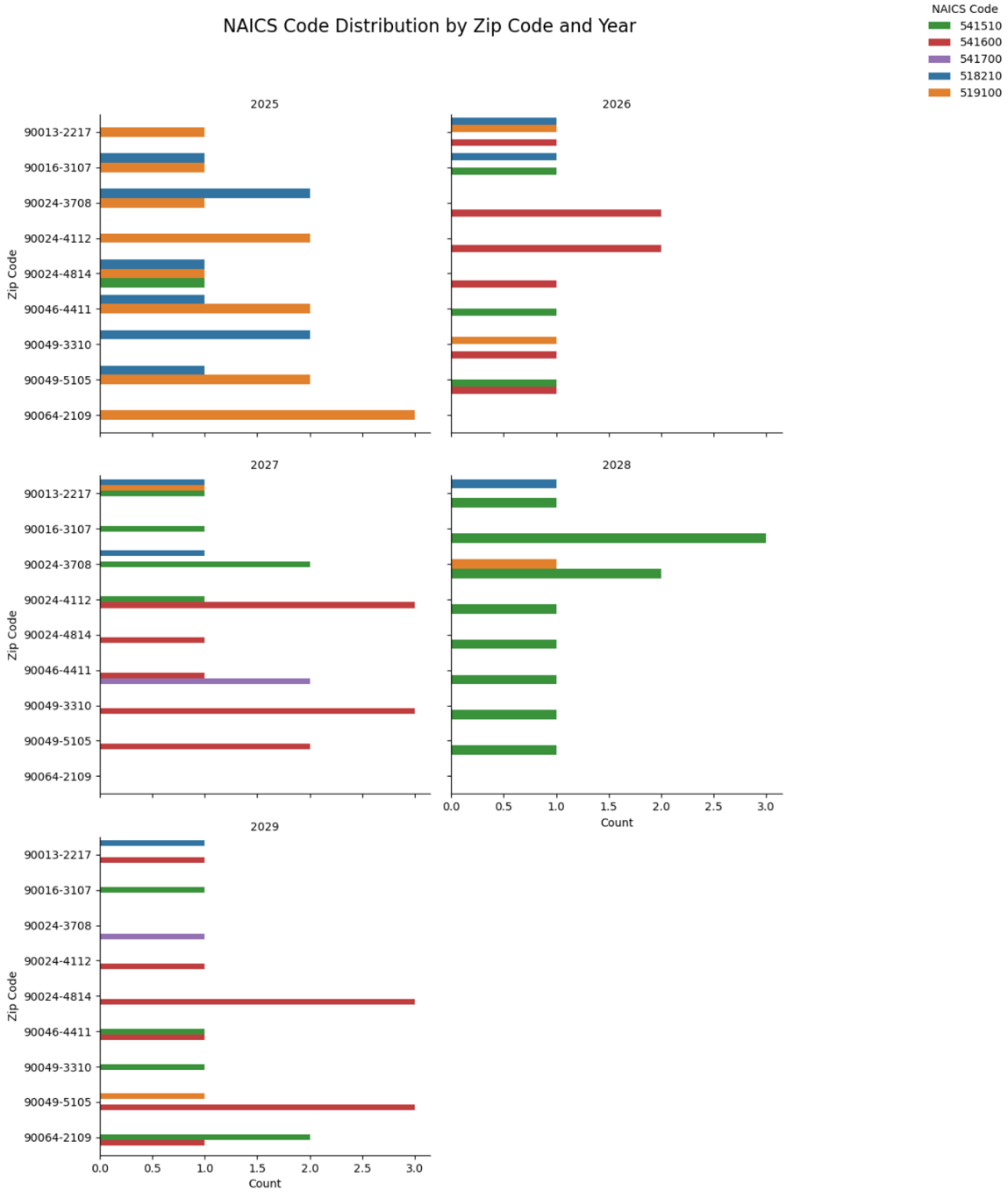
The final model achieved an accuracy of 80%, with a precision of 89% and a recall of 85%. The Extra Trees Classifier provided robust predictions, outperforming other models tested. Cross-validation did not quite work well enough, however, indicating that we would do well to carry out further testing on other zip codes to prevent overfitting.

Key Metrics:

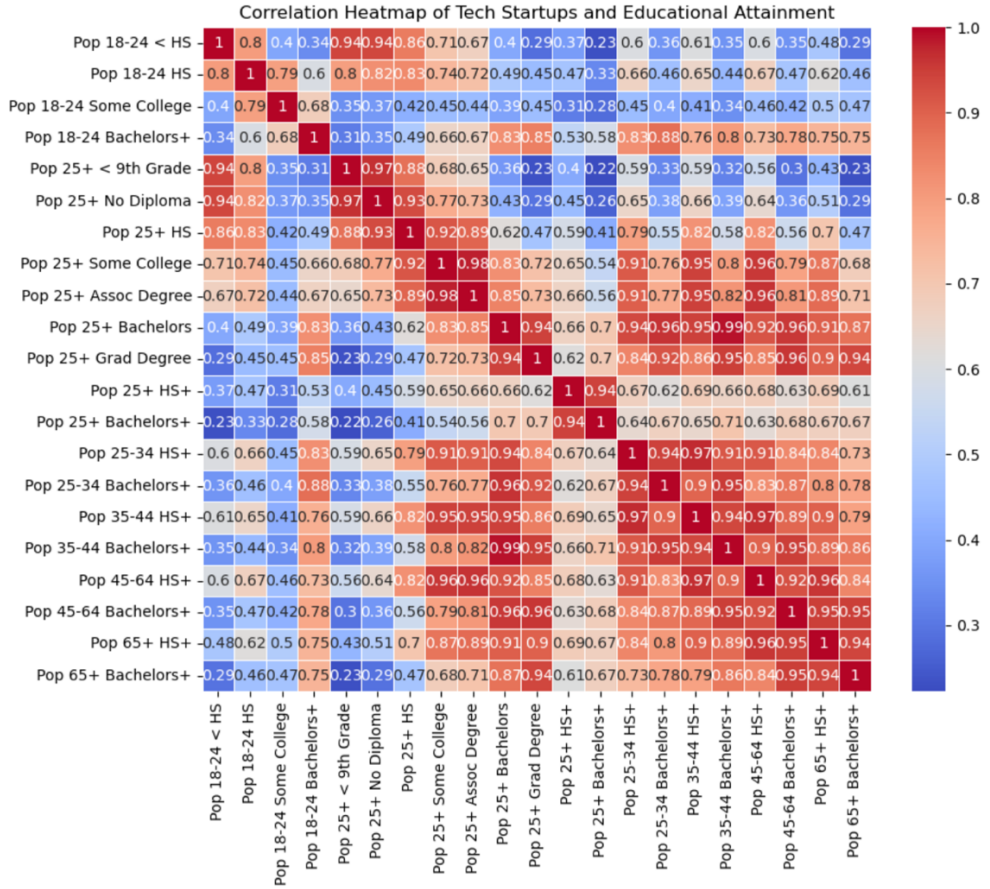
- **Accuracy**: 80%
- **Precision**: 89%
- **Recall**: 85%

- **Visualizations:**

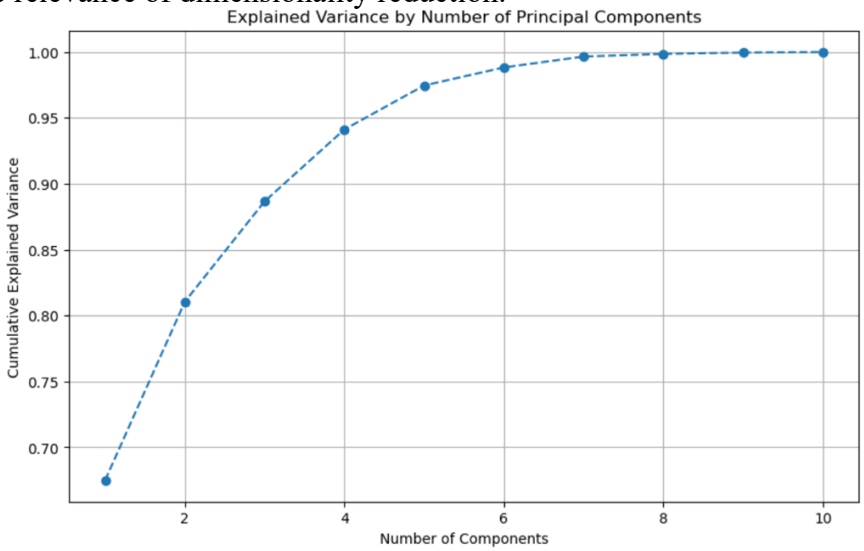
  Facet Grid for Predicted NAICS Startups (2025-2029): A facet grid showing the predicted NAICS codes for each year by zip code.



- **Heatmap of Educational Attainment vs. Tech Startups**: Demonstrates a strong positive correlation between areas with a higher percentage of bachelor's degrees and tech startups.

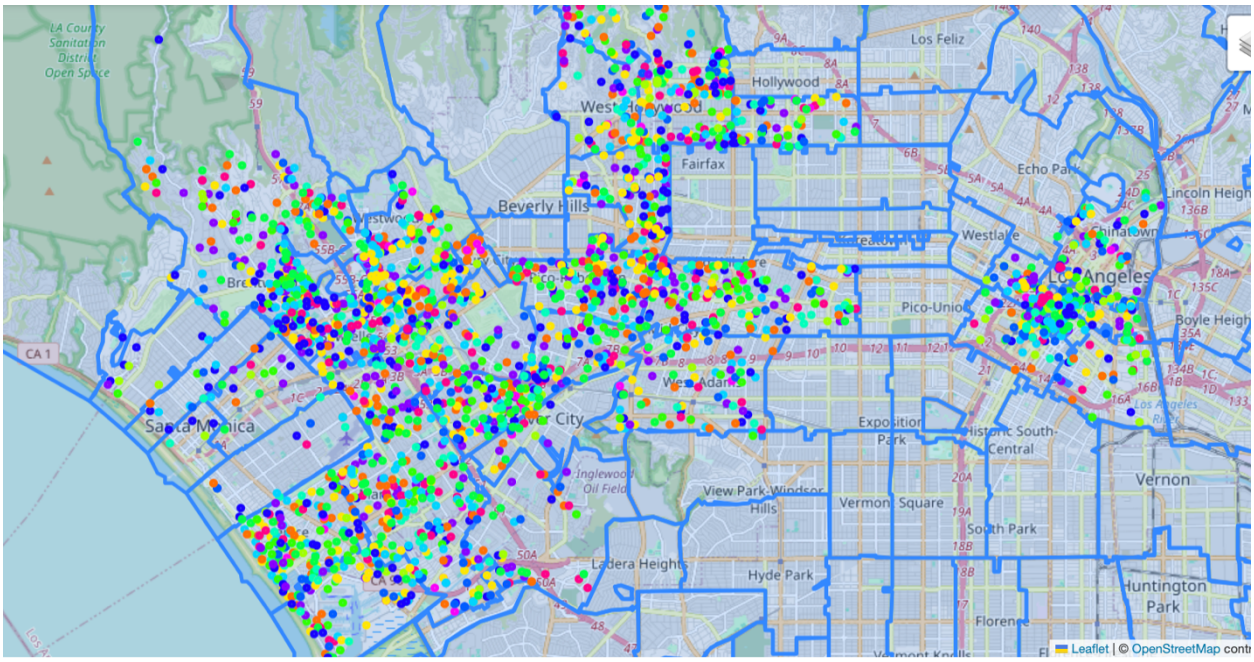


- **Principal Component Variance Chart**: Shows that PC1 captures 75.7% of the variance, confirming the relevance of dimensionality reduction.




```
Explained Variance by Principal Component:
     Explained Variance
PC1            0.675004
PC2            0.135521
PC3            0.076165
PC4            0.054548
PC5            0.033364
```

- **Clustering Predictions Map**: Displays the predicted clustering of tech startups over the next five years, color-coded by year.



---

5. Key Findings

- Educational attainment has a strong influence on tech startup clustering. In particular, areas such as Santa Monica, Venice, and Downtown Los Angeles showed a high correlation between the percentage of bachelor's degree holders and the number of tech startups.

- Tech startup growth is expected to continue in these areas over the next eight years, with a predicted increase in NAICS codes related to technology sectors.

---

6. Recommendations

Based on the findings, I recommend the following:

- **Increase investment in educational infrastructure** in areas like Santa Monica and Venice to further foster the growth of tech startups.

- **Develop targeted urban development policies** that align educational opportunities with business incentives, particularly in areas with high potential for tech growth.

- **Monitor changes in tech startup clustering** over time to adjust policies dynamically, ensuring continued support for the tech sector in key areas.

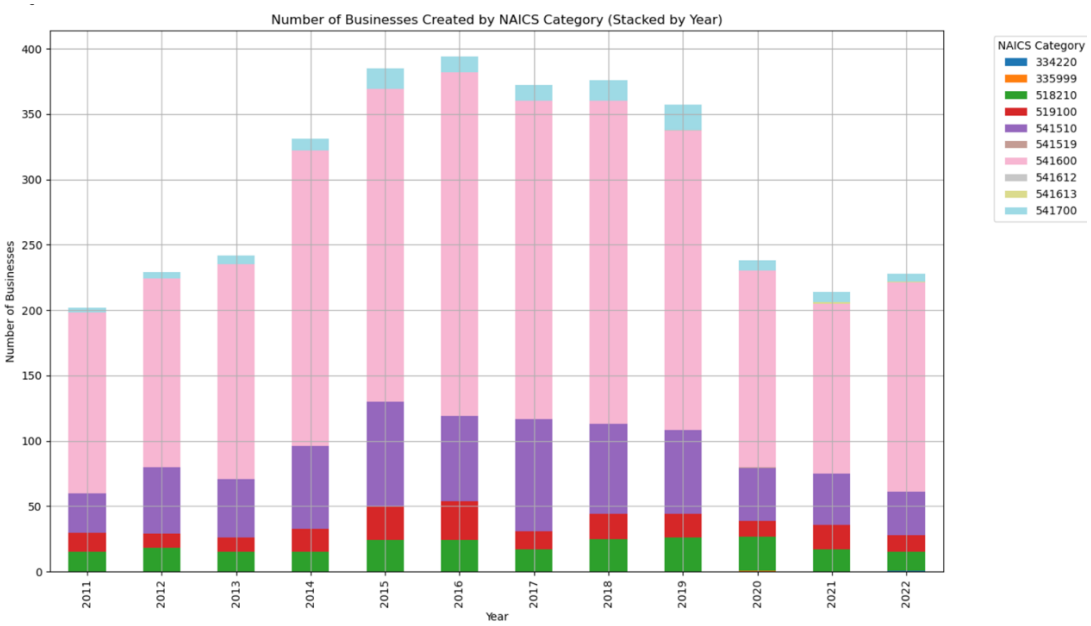
---

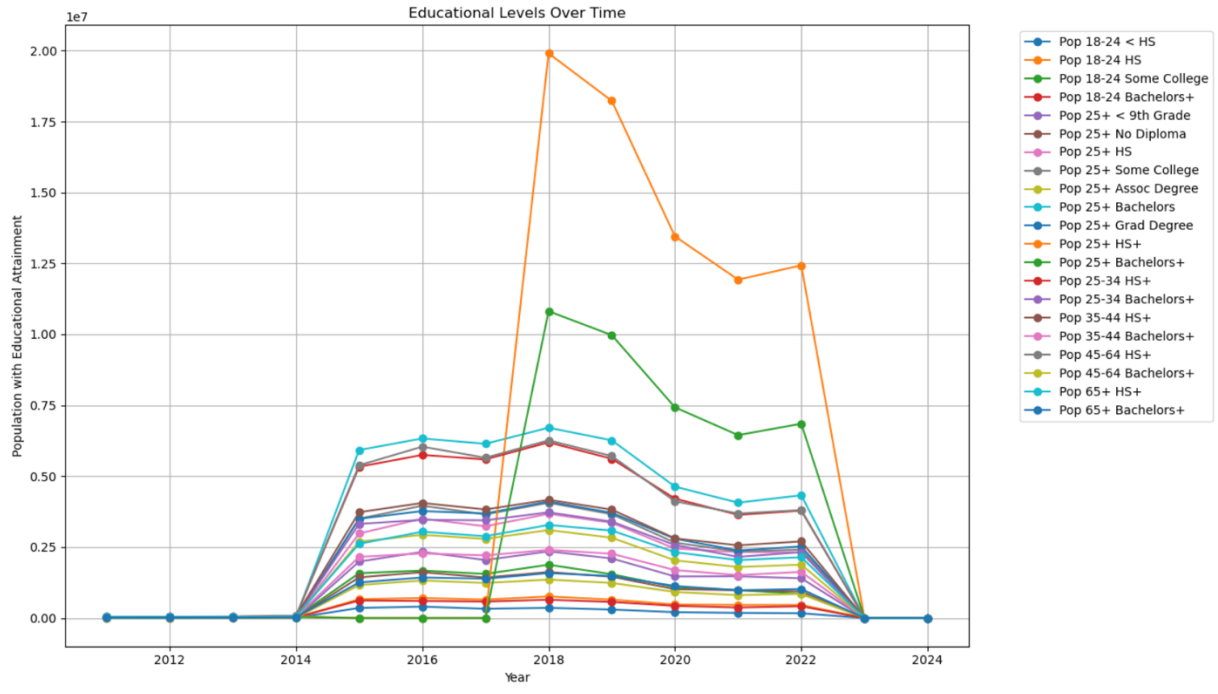
7. Conclusion

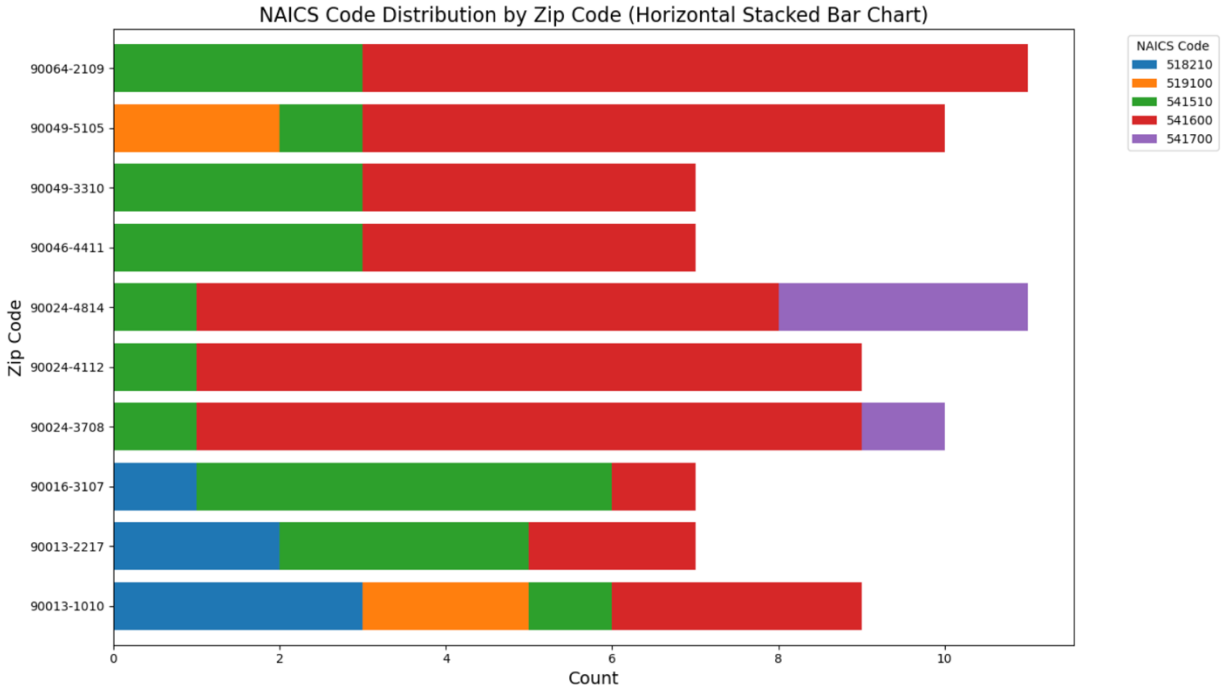
This project successfully predicted the future clustering of tech startups in Los Angeles, with the Extra Trees Classifier providing accurate predictions. The correlation between educational attainment and tech startup growth emphasizes the importance of educational investments.

Future work could include incorporating more recent data (post-2024) and exploring additional models to improve prediction accuracy.

Appendix






NAICS Code Distribution by Zip Code (Horizontal Stacked Bar Chart)
Zip Code
90064-2109
90049-5105
90049-3310
90046-4411
90024-4814
90024-4112
90024-3708
90016-3107
90013-2217
90013-1010
0
2
4
6
8
10
Count
NAICS Code
518210
519100
541510
541600
541700