BSAN6050

Case Study: LA Business Profitability

Analysis of Business Data And Predictors of Sales Growth

Rachel Pham

Herat Devisha

Diego Estuar

Yannick Angouo Lopes

Andrew Morris

Riley Nickel

Table of Contents

Executive Summary

The success of businesses relies heavily on sustained revenue, yet predicting this metric remains a challenge. Recognizing this, our study aimed to optimize sales growth potential by focusing on specific factors within the LA business landscape. Leveraging comprehensive datasets from the National Establishment Time-Series (NETS) database and the US Census Bureau, we conducted extensive analysis and built predictive models to address two critical research questions: "For a given business type and target customer, what location (ZIP code) will result in the highest sales growth?" and "Using the data available, is it possible to predict sales growth of a business based on the demographic and market environment?" Using cluster analysis, we segmented the market into distinct customer groups based on demographic factors. This approach allowed us to identify six unique customer clusters within Los Angeles, ranging from "Affluent and Educated" to "Working-class, Hispanic Dominant". Subsequently, we conducted linear regression to determine the best location (ZIP code) for a specific business type targeting each cluster. The process involved predicting sales growth for multiple ZIP codes and ranking them to recommend the most promising locations for businesses catering to each cluster. Additionally, in pursuit of predictive analytics, we developed a Random Forest classification model. This model, created on a 50,000 record sample of our full business and demographic dataset, demonstrated promising outcomes, achieving a 93% accuracy rate. The model effectively predicted sales growth, providing valuable insights for businesses planning to establish operations in the Los Angeles area.

Introduction and Research Question

For new businesses, achieving and sustaining profitable growth over time is challenging. Over 20 percent of all new businesses fail within the first year and over 65 percent are forced to close their doors within the first 10 years of operation (Statista, 2023). With new business owners facing tremendous uncertainty around the future success of their companies, profitability is an important metric and should be at the forefront for business related decision-making. Although profit is commonly forecasted within businesses, it is difficult to predict accurately. Strategy built around inaccurate predictions can lead to

poorly-informed investments, negatively impact financial outcomes, and limit the potential for future success. Part of the problem of forecasting profitability is that there are a number of factors that play a determining role for any given business such as focal industry, location (ZIP code), sales growth, and demographic factors (Cegiełka, 2023). Out of the variables present in the data available to us, we hypothesized that sales growth and profitability would have the strongest positive correlation, highest predictive power and greatest potential for accurate analysis. Based on this hypothesis, we aim to answer the following research question: *For a given business type and target customer, what location (ZIP code) will result in the highest sales growth?*

Because sales growth and profitability around new businesses and startups boasts important research potential, the subject of this analysis is not dissimilar to prior-research. Although these questions have been well-researched in the past, there are few manuscripts that are easily interpretable by potential new business owners in any industry and provide direct, actionable insights. With these goals in mind, this paper will highlight the process and results of our analysis, as well as determine the best possible location for a new business for any industry and deliver a predictive model that will return predicted sales growth from a number of business and demographic variables.

Data Processing

In our analysis, the initial dataset was derived from the National Establishment Time-Series (NETS) database. This dataset comprised over 600,000 data points collected in 2020, providing detailed information on individual businesses in the Los Angeles area. Key variables in the NETS data included metrics such as the previous year's sales, the quartile of sales growth, demographics of the owner, and business type, offering a comprehensive view of the business landscape in Los Angeles. The second dataset, obtained from the US Census Bureau, encompassed demographic data at the ZIP code level for California, spanning the years 2011 to 2020. To align with the objectives of cluster analysis and regression, our team only retained the 2020 data from the Census dataset.

Following this, a left merge was executed, with the NETS data serving as the left dataset, ensuring the preservation of each business's information in the final merged dataset. Subsequent data cleaning steps involved the removal of non-2020 years from the Census data and the exclusion of rows where the 'Relocated' variable equaled 1, emphasizing a focus on businesses that had maintained their location. These preprocessing steps provided a foundation for a targeted examination of factors influencing businesses in Los Angeles, integrating both establishment-specific details and broader demographic insights.

Methodology

In this section, we present the method we conducted to answer the above research question. Overall, we perform cluster analysis and linear regression to determine the best location to open a business given the business type and a target customer group.

1. Cluster analysis

A growing and sustainable business understands its customers. Therefore, we perform cluster analysis to cluster the location census data into several groups, where different locations within one group share the same demographic characteristics. This segments the market to smaller and specialized target customer groups, helping owners determine which locations best suit their future businesses. The following shows how we perform cluster analysis step-by-step:

a. Load the location census data and choose a subset of columns for clustering

In this project, we decided to use the following 8 columns: pop_black, pop_asian, pop_white_all, pop_hispanic, median_age, per_capita_income, edu_highschool_percent, edu_bachelor_percent, which encapsulate multiple aspects of the demographic information.

```
# Load the dataset
file_path = 'location_census_CA.csv'  # Replace with your file path
census_data = pd.read_csv(file_path)
census_data = census_data[census_data['year'] == YEAR_TO_USE]

# Specify the columns to be used for clustering
required_columns = [
    'pop_black', 'pop_asian', 'pop_white_all', 'pop_hispanic',
    'median_age', 'per_capita_income', 'edu_highschool_percent', 'edu_bachelor_percent'
]
```

b. Data processing

We first fill in the missing cells using the median value of their corresponding columns. After that, we standardize each column to have 0 mean and 1 standard deviation. The standardization process makes sure different columns have a similar range of values, avoiding one column to dominate the others in clustering.

```
# Check for missing values and fill them with the median if any
if census_data[required_columns].isnull().sum().sum() > 0:
    census_data[required_columns] = census_data[required_columns].fillna(census_data[required_columns].median())

# Normalize the data
scaler = StandardScaler()
scaled_data = scaler.fit_transform(census_data[required_columns])
```

c. Run clustering

We then run K-Means algorithm to cluster the location data into 6 clusters. We played around with the number of clusters and found 6 to work the best, i.e., the clusters have the most meaningful interpretation.

```
# Clustering the data into 6 clusters
kmeans_6 = KMeans(n_clusters=6, init='k-means++', max_iter=300, n_init=10, random_state=0)
census_data['Cluster_6'] = kmeans_6.fit_predict(scaled_data)
```

d. Save cluster data to files

We save each cluster data into a separate .csv file for easier further processing and analysis.

```
# Saving clusters to different CSV files
output_columns = ['GEOID'] + required_columns + ['Cluster_6']
for cluster_id in range(6):
    cluster_data = census_data[census_data['Cluster_6'] == cluster_id]
    cluster_data.to_csv(f'cluster_{cluster_id}.csv', index=False, columns=output_columns)
```

Each file will look like the following:

```
cluster_data[0]
✓ 0.0s
```

	GEOID	pop_black	pop_asian	pop_white_all
0	90010	270	2438	951
1	90035	1786	1849	21093
2	90036	2187	7744	24163
3	90048	978	1700	17059
4	90049	449	2973	30175
...	...	...	...	...
253	96140	12	68	1052
254	96141	0	22	441
255	96145	78	65	2442
256	96146	0	0	1030
257	96161	63	272	16976

258 rows × 10 columns

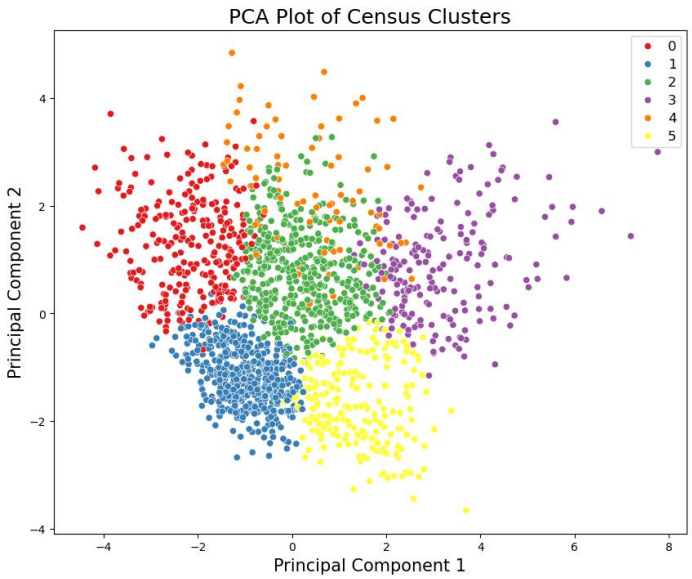
e. Cluster visualization

We use PCA to reduce the dimension of the data from 8 to 2 and visualize the 6 clusters on a 2-D map.

```
# Perform PCA for visualization
pca = PCA(n_components=2)
principal_components = pca.fit_transform(scaled_data)
pc_df = pd.DataFrame(data=principal_components, columns=['PC1', 'PC2'])
pc_df['Cluster_6'] = census_data['Cluster_6'].values

# Plotting the clusters using PCA components
plt.figure(figsize=(10, 8))
sns.scatterplot(x="PC1", y="PC2", hue="Cluster_6", palette="Set1", data=pc_df, legend="full")
plt.title('PCA Plot of Census Clusters', fontsize=18)
plt.xlabel('Principal Component 1', fontsize=15)
plt.ylabel('Principal Component 2', fontsize=15)
# change font size of legend, code below
legend = plt.legend()
plt.setp(legend.get_texts(), fontsize='12')
plt.savefig('pca_plot.png', bbox_inches='tight')
plt.close()
```

Each color in the visualization denotes a different cluster. The visualization shows that the 6 clusters are well distinguished, while different locations within one cluster are close. This verifies that the clustering algorithm is able to cluster the location data into multiple groups that share demographic characteristics.



f. Cluster interpretation

The final step of cluster analysis is to interpret the clusters to have a better understanding of what customer group each cluster represents. To do this, we look closely at the saved cluster data files

and interpret each cluster based on its demographic information. We summarize the interpretation as follows:

- Cluster 0: Affluent and Educated

  This cluster, with 258 entries, is characterized by being older with a median age around 46, a high per capita income, and high levels of high school and bachelor's degree attainment. The population is predominantly White, with very low counts of Black and Asian populations, and a low Hispanic population.

- Cluster 1: Mid-Income, Diverse Population

  With 552 entries, this cluster has an older population with a median age around 49. It has a lower per capita income and lower levels of educational attainment compared to Cluster 0. The population is predominantly White, with significant counts of other races.

- Cluster 2: Young and Multicultural

  This cluster has 442 entries, featuring a relatively younger demographic with a median age in the late 30s, and a moderate per capita income. Educational attainment is moderate, with a significant percentage of bachelor's degree holders. The population is predominantly White, with a diverse mix of other races.

- Cluster 3: Working-Class, Hispanic Dominant

  Comprising 195 entries, this cluster has the youngest median age around 33, indicating a younger demographic. It has the lowest per capita income among the clusters, and the education levels are also the lowest, with the lowest percentages of high school and bachelor's degree holders. The population is a mix, with a dominant Hispanic presence and moderate counts of other races.

- Cluster 4: Middle-Aged, Diverse Middle Class

  With 83 entries, this cluster is characterized by a middle-aged demographic with a median age around 39, and a middle-income level. The education levels are higher than Clusters 1

and 3. The population is predominantly Asian, with significant counts of White and other races.

- Cluster 5: Middle-Aged, Diverse Middle Class

  This cluster, with 234 entries, has a relatively young median age around 33 and the lowest per capita income. Educational attainment is the lowest among the clusters. The population is predominantly Hispanic, with a mix of other races present.

2. Linear regression for choosing the best location (zip code)

Clustering determines different customer groups an owner can target for their new business. To further narrow down to one specific location that potentially yields the highest sales growth, we additionally perform linear regression on top of the clusters. We describe each step in detail below:

a. Independent and dependent variables

We choose 4 business-related columns, 'PubPriv', 'EmpHereC', 'NAICS', 'Name_4', and the 8 demographic variables described above as the independent variables to predict the dependent variable ‘SalesGrowth'.

```
YEAR_TO_USE = 2020
COLUMNS_TO_USE = [
    # business related, can include more if needed
    'PubPriv', 'EmpHereC', 'NAICS', 'Name_4',
    # demographic related
    'pop_black', 'pop_asian', 'pop_white_all', 'pop_hispanic',
    'median_age', 'per_capita_income', 'edu_highschool_percent', 'edu_bachelor_percent'
]
DEPENDENT_COLUMN = ['SalesGrowth']
✓ 0.0s
```

b. Merge business data and location data

```
survival_data = pd.read_csv('Survival_LA_City.csv')
location_data = pd.read_csv('location_census_CA.csv')
location_data = location_data[location_data['year'] == YEAR_TO_USE]
location_data = location_data.rename(columns={'GEOID': 'ZipCode'})
```

```
data = pd.merge(survival_data, location_data, on='ZipCode', how='left')
data = data[COLUMNS_TO_USE + DEPENDENT_COLUMN]
data = data.dropna()
data = pd.get_dummies(data, drop_first=True)
```

c. Run the linear regression model

```
X = np.asarray(data.drop(DEPENDENT_COLUMN, axis=1)).astype(np.float64)
Y = np.asarray(data[DEPENDENT_COLUMN].values.squeeze()).astype(np.float64)

model = sm.OLS(Y, X)
results = model.fit()
print(results.summary())
```

d. Linear regression results

```
                                 OLS Regression Results
================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.863
Model:                            OLS   Adj. R-squared (uncentered):              0.863
Method:                 Least Squares   F-statistic:                          1.474e+04
Date:                Wed, 13 Dec 2023   Prob (F-statistic):                        0.00
Time:                        17:25:25   Log-Likelihood:                     -7.8339e+05
No. Observations:              615555   AIC:                                  1.567e+06
Df Residuals:                  615292   BIC:                                  1.570e+06
Df Model:                         263
Covariance Type:            nonrobust
```

e. Predict sales growth for a new business and determine the best location (zip code)

After we build the linear model, we can use it to make sales growth predictions for a new business. To do this, we first determine the business-related information, including the values for PubPriv, EmpHereC, NAICS, and Name_4 columns.

```
new_business = {
    'PubPriv': 'Y',
    'EmpHereC': 2,
    'NAICS': 621111,
    'Name_4': 'Restaurants and Other Eating Places',
    'pop_black': None,
    'pop_asian': None,
    'pop_white_all': None,
    'pop_hispanic': None,
    'median_age': None,
    'per_capita_income': None,
    'edu_highschool_percent': None,
    'edu_bachelor_percent': None
}
✓ 0.0s
```

To determine what is the best location to open this business for each customer group, we loop through the 6 clusters, and for each zip code in each cluster, we:

- Take the demographic information of that zip code and include it as the demographic-related independent variables
- Input this to the linear to model to obtain the sales growth prediction
- Repeat for each zip code in the cluster
- After obtaining the predictions for all zip codes, we rank different zip codes inside a cluster based on their predictions. For example, for cluster 0, if the zip code 90024 is predicted to have the highest sales growth then it means the owner should open the business in this zip code (if he/she targets customers in cluster 0).

The Python code to perform this step is as follows:


```
cluster_prediction = {
    'cluster_0': {},
    'cluster_1': {},
    'cluster_2': {},
    'cluster_3': {},
    'cluster_4': {},
    'cluster_5': {},
}

for i, cluster in enumerate(cluster_data):
    # loop through each zip code in cluster
    # and extract the demographic data
    for index, row in cluster.iterrows():
        new_business_zip_code = deepcopy(new_business)
        new_business_zip_code['pop_black'] = row['pop_black']
        new_business_zip_code['pop_asian'] = row['pop_asian']
        new_business_zip_code['pop_white_all'] = row['pop_white_all']
        new_business_zip_code['pop_hispanic'] = row['pop_hispanic']
        new_business_zip_code['median_age'] = row['median_age']
        new_business_zip_code['per_capita_income'] = row['per_capita_income']
        new_business_zip_code['edu_highschool_percent'] = row['edu_highschool_percent']
        new_business_zip_code['edu_bachelor_percent'] = row['edu_bachelor_percent']
        new_business_zip_code = pd.DataFrame([new_business_zip_code])
        new_business_with_dummy = get_dummy_df(new_business_zip_code, data.columns)

        # make prediction from the linear model above
        new_business_with_dummy_arr = np.asanyarray(new_business_with_dummy.drop(DEPENDENT_COLUMN, axis=1)).astype(np.float64)
        prediction = results.predict(new_business_with_dummy_arr)

        # save the prediction
        cluster_prediction['cluster_' + str(i)][row['GEOID']] = prediction[0]
```

We then save the predictions for all zip codes within a cluster to a json file, sorted based on the sales growth prediction. The json file looks like the following:

This result indicates that 95595 is the best zip code to open the business if the owner wants to target the customer group in cluster 0, or in other words Affluent and Educated customers. On the other hand, if the owner wants to target Mid-Income, Diverse Population customers (cluster 1), he/she should open the business in zip code 96136.

Conclusion and Recommendations

Using the datasets from the National Establishment Time-Series (NETS) database and the US Census Bureau, we conducted extensive analysis and built predictive models to address two critical research questions: "For a given business type and target customer, what location (ZIP code) will result in the highest sales growth?" and "Using the data available, is it possible to predict sales growth of a business based on the demographic and market environment?" Using cluster analysis, we segmented the market into distinct customer groups based on demographic factors. This approach allowed us to identify six unique customer clusters within Los Angeles, ranging from "Affluent and Educated" to "Working-class, Hispanic Dominant". Subsequently, we conducted linear regression to determine the best location (ZIP code) for a specific business type targeting each cluster. The process involved predicting sales growth for multiple ZIP codes and ranking them to recommend the most promising locations for businesses catering to each cluster. Additionally, in pursuit of predictive analytics, we developed a Random Forest classification model. This model, created on a 50,000 record sample of our full business and demographic dataset, demonstrated

promising outcomes, achieving a 93% accuracy rate. The model effectively predicted sales growth, providing valuable insights for businesses planning to establish operations in the Los Angeles area.

Through this analysis, we recommend that future business owners use this cluster analysis, or similar models, to determine the best possible location for a new business for any industry before choosing a location. Additionally, the use of a predictive model to return predicted sales growth from a number of business and demographic variables can help business owners predict future sales growth and adjust their business strategy accordingly. Future research should look to expand on this subject by including additional relevant business variables, analyzing the clusters and linear regression for each business type, and applying the predictive model to a larger or more robust dataset to improve accuracy.

Appendix

Cegiełka, M. (2020, January 1). Factors determining the survival of new companies. Central European Economic Journal. https://sciendo.com/article/10.2478/ceej-2020-0021

Statista.com (2023, November 30). New Business Success Rate U.S. 2022. Statista. https://www.statista.com/statistics/725044/survival-rate-new-business-united-states/