Accenture Office Location Selection

AI Studio Final Presentation

Accenture - Los Angeles
December 14, 2022

Introductions

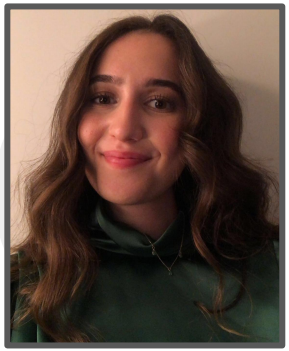
Meet Our Team!

Anni Bamwenda

Whittier College ‘22, majoring in Math and Computer Science

Hritika Chaturvedi

UCLA ’25, majoring in Computational and Systems Biology



Gila Kohanbash

USC ‘24, majoring in Computer Science and Computer Engineering

Samantha Preciado

Cal State Long Beach ’25, majoring in Computer Science

Our AI Studio TA and Challenge Advisors

Shruthi Srinarasi
AI Studio TA

Jodi Yip
Challenge Advisor

Timo Budiono
Challenge Advisor

Presentation Agenda

1. Project Overview
2. Data Preprocessing
3. Model Selection and Evaluation
4. Final Thoughts

AI Studio Project Overview

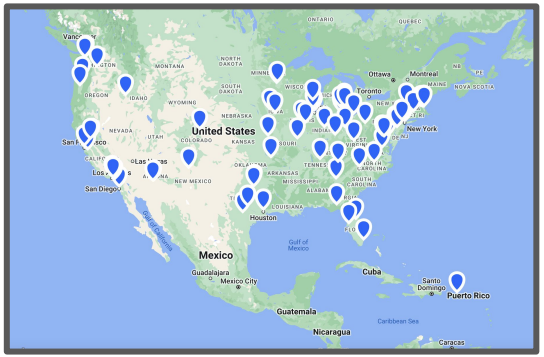
Predict the Optimal Location for a Future Accenture Office.

Aggregate and clean provided datasets, utilize a machine learning algorithm, and propose at least 3 cities where Accenture should open a new location. Using US cities economic data and Fortune 500 Corporate data.

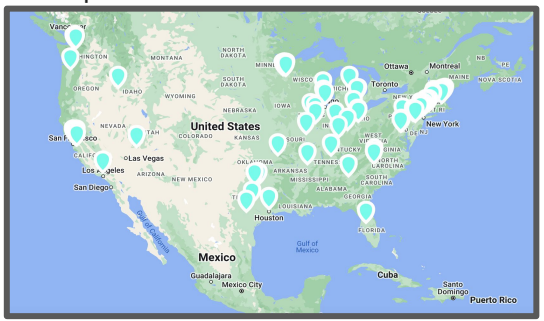
Given Datasets

- Accenture Offices All (from Accenture): CSV file of all the Accenture offices in North America separated by region (West, Midwest, Northeast, South)
- U.S. Cities (from Census): CSV file of major US Cities including population, latitude and longitude, etc.
- Fortune 500 Corporate Headquarters (from Census): CSV file containing the ranking of companies, number of employees, revenue, profit margins, etc.

Accenture Office Locations



Top 100 Fortune 500 in the US

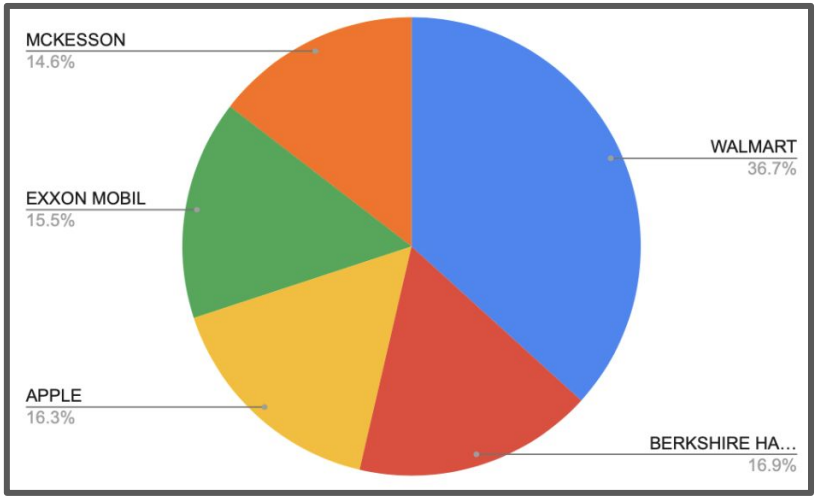


Our Goal

1. Data manipulation
2. Using machine learning to predict new office locations via these factors:
    a. Population density
    b. No. of fortune 500 companies in a city
    c. Population of the city
3. Identify our model's success using 2 evaluation metrics namely accuracy score and log loss

Business Consideration – Purpose

Beneficial to Accenture in that it offers exposure to new clients, maintaining long-lasting relationships with existing clients, and maximizing their profits while minimizing loss



Revenue of the top 5 Fortune 500 companies; under the assumption of Accenture's clients
Source: given dataset

Our Approach

End of Oct.

Data Cleaning

Drop columns from U.S. cities dataset that are not helpful in solving our business problem (ex. city ascii, state ID, country flips, military and timezone); repeating values.

Nov.

Feature Engineering

Introducing 3 new features to the U.S. cities dataset to allow classification when predicting.

Mid Nov.

Machine Learning

Use clean data to train a machine learning model and make predictions. Use logistic regression for binary classification.

Dec.

K-Means Clustering (in progress)

Unsupervised machine learning algorithm used to group items using a distance algorithm; calculates how close data points are to one another.

Resources We Leveraged

ATLIST

Data Preprocessing

Data Cleaning

- Drop repeated columns and unnecessary columns from the datasets that are not helpful in solving our business problem (ex. city ascii, state ID, country flips, military and timezone
- Drop rows with missing values.
- Drop cities with 0 Fortune 500 companies
- Create a new dataset (dataframe) from the existing datasets to use for modeling

Code

Code snippet used to drop cities with 0 fortune 500 company offices

```
df.drop(df[(df['no_of_fortune_500'] == 0)].index, inplace=True)
```

Data shape before cleaning:

```
Dataframe shape: (30409, 9)
```

Data shape after cleaning:

```
Dataframe shape: (209, 8)
```

Feature Engineering

Introducing 2 new features to the U.S. cities dataset to allow classification when predicting.

- Binary (1/0) labels to each of the cities
  - 1: Accenture has an office
  - 0: Accenture doesn't have an office
- No. of fortune 500 company in a city

Source: Will Koehrsen, Feature Engineering: What Powers Machine Learning, Medium,, accessed 23 September 2022, https://towardsdatascience.com/feature-engineering-what-powers-machine-learning-93ab191bcc2d>

Code

```
# Feature engineering
# changing the 'flag' to 1 for locations where accenture has offices
for city in df_acc_cities.City:
  df_cities.loc[df_cities.city_name == city,'flag'] = 1

# counting the number of fortune 500 companies in a city then changing the value to the count in our '
# create a dictionary for the df_fortune.CITY with cities as keys and frequency/count as the value
fortune500_dict = {}

for row in df_fortune.CITY:
  if row in fortune500_dict:
    fortune500_dict[row] += 1

  else:
    fortune500_dict[row] = 1

# print(fortune500_dict)

# now we put the values of the dictionary keys into our dataframe
list_cities = list(df_cities.city_name)
for city in list_cities:
  if city.upper() in fortune500_dict:
    df_cities.at[list_cities.index(city), 'no_of_fortune_500'] = fortune500_dict[city.upper()]
```

Code

Dataframe after feature engineering

	city_name	state_name	population	density	flag	no_of_fortune_500
0	New York	New York	18680025	10768.0	1	39
1	Los Angeles	California	12531334	3267.0	1	3
2	Chicago	Illinois	8586888	4576.0	1	8
3	Miami	Florida	6076316	4945.0	1	3
4	Dallas	Texas	5910669	1522.0	1	7
5	Houston	Texas	5724418	1394.0	1	10

Splitting the data

We split the dataset into training and test sets.

- Training set: to train our machine learning model. Results will be used on the test set. The training size was 50%
- Test set: Use results from training set to predict new office locations (the label). The test size was 50%

Model Selection and Evaluation

Machine Learning

Introduction

Use clean data to train a machine learning model and make predictions.

Use logistic regression for binary classification.

- Used to solve classification problems. Ex. email apps classify emails as spam or not spam
  - Use existing data to learn what kind of emails have been previously classified to be spam or not then use it to predict future emails as spam or not spam

Model Comparison

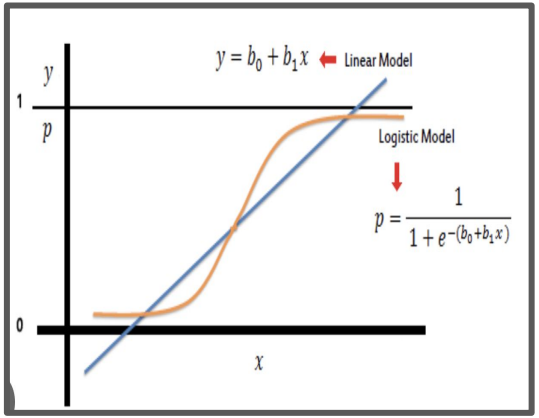
Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.

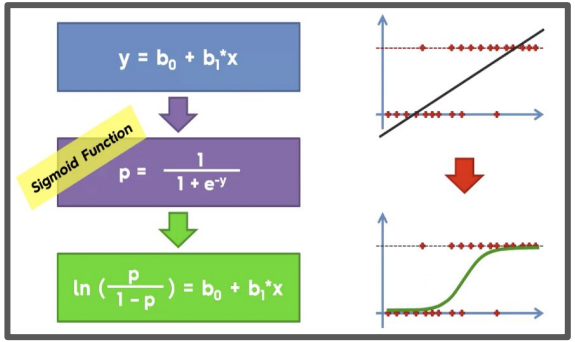
Machine Learning

Applying it to our business problem.

We'll use binary classification to predict whether Accenture should open a new office location (1) or not (0).

Logistic regression model will learn from our training set, then hypertune/modify the features (columns) to give us better predictions.





Source: ResearchGate.net

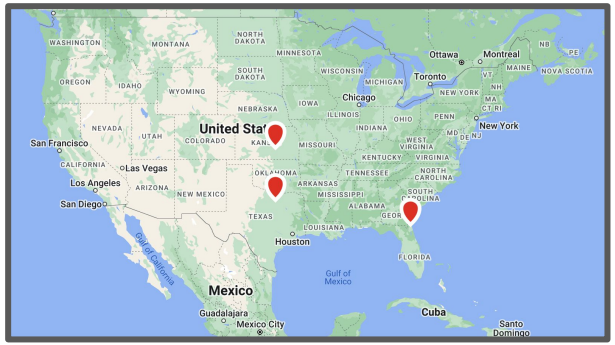
Code

```
'''
Fit a Logistic Regression classifier to the training data X_train, y_train.
Return the loss and accuracy of resulting predictions on the test set.
Parameters:
    C = Factor that controls how much regularization is applied to the model.
'''
  # 1. Create the  scikit-learn LogisticRegression model object below and assign to variable 'model'
model = sklearn.linear_model.LogisticRegression()

# 2. Fit the model to the training data below

model.fit(X_train, y_train)

# 3. Make predictions on the test data using the predict_proba() method and assign the result to the
# variable 'probability_predictions' below


probability_predictions = model.predict_proba(X_test)

# 4. Compute the log loss on 'probability_predictions' and save the result to the variable 'l_loss' below
l_loss = log_loss(y_test, probability_predictions)

# 5. Make predictions on the test data using the predict() method and assign the result to the
# variable 'class_label_predictions' below
class_label_predictions = model.predict(X_test)

# 6. Compute the accuracy score on 'class_label_predictions' and save the result to the variable 'acc_score' below
acc_score = accuracy_score(y_test, class_label_predictions)
```

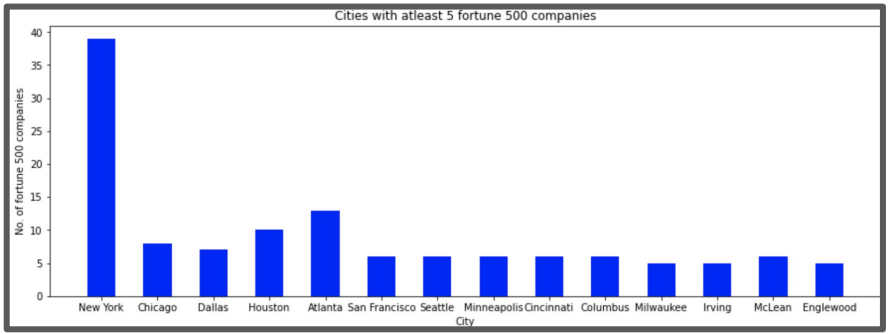
Insights and Key Findings

Our machine learning model predicted 3 cities where Accenture should open new office locations namely:

- Jacksonville, Florida
- Fort Worth, Texas
- Wichita, Kansas



We found several cities that had at least 5 fortune 500 companies in their location



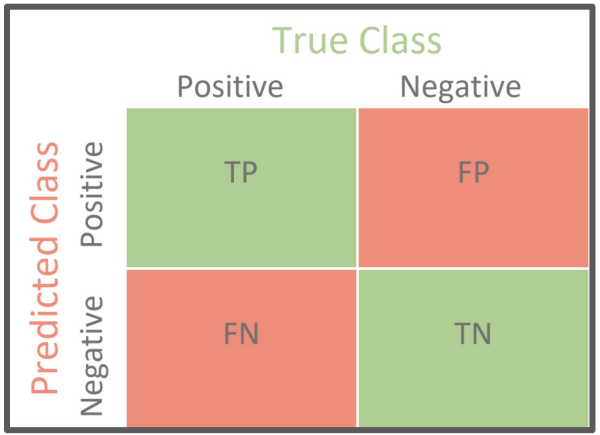
Model Accuracy

We used accuracy score and log loss to evaluate our machine learning model.

Log-loss is indicative of how close the prediction probability is to the corresponding actual/true value. The more the predicted probability diverges from the actual value, the higher is the log loss value. Our model's log-loss is 47%.

Accuracy score is the proportion of correct predictions over the total predictions. Our model's accuracy score is 86%.

```
Log loss: 0.4783573133311523
Accuracy Score: 0.8666666666666667
```



Final Thoughts

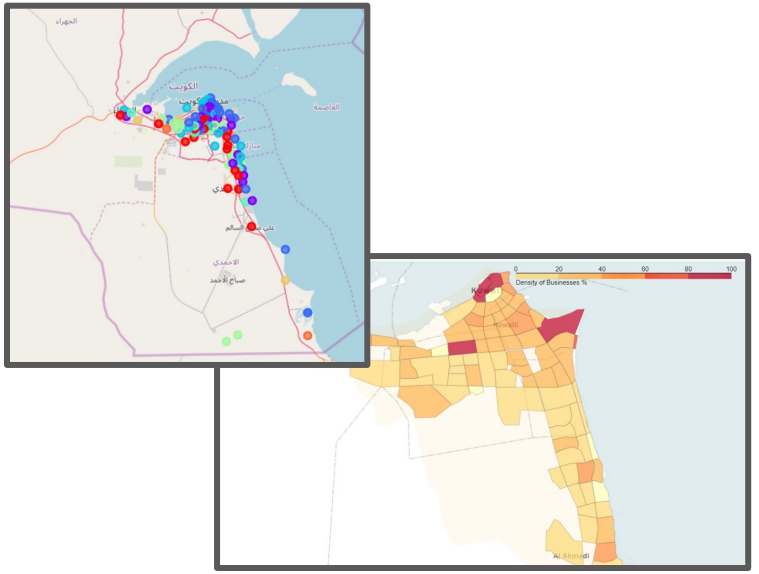
Potential Next Steps

- Include companies outside the metropolitan city as part of the main city count
- Include sum of revenue for fortune 500 companies in each city
- K-mean clustering
- Ethical considerations when pinpointing office locations

K-Means Clustering

K-means Clustering is an unsupervised machine learning algorithm used to group items using a distance algorithm; calculates how close data points are to one another.

Also consider how close the office would be to airport(s), hotels, median income of the area, and whether the location is downtown/uptown.



Source: Frith J 1968, Using Machine Learning to Identify the Best Location for your Business, LinkedIn, accessed 23 September 2022, <https://www.linkedin.com/pulse/using-machine-learning-identify-best-area-kuwait-your-dawoud/>


Questions?