Enhancing Nonprofit Organization Outreach: Leveraging Predictive Analytics for Program Expansion and CheckIn Projection

Mauren Shih (mshih5@uic.edu), Ying-Hsin Lai (ylai27@uic.edu), Xin Yie Ng (xng4@uic.edu), Samridhi Madan (smadan7@uic.edu), Khwaja Mominuddin (kmomin3@uic.edu),
Professor Fatemeh Sarayloo (fsaraylo@uic.edu), Professor Kimberly VanderSchaaf (kvanders@uic.edu)

UIC BUSINESS UNO United Neighborhood Organization — informs ANALYTICS CONFERENCE


Mauren Shih


Ying-Hsin Lai


Xin Yie Ng


Samridhi Madan


Khwaja Mominuddin

Abstract

This study applies advanced machine learning techniques to analyze participant check-ins for the Carrera 5K Race and explore opportunities for expanding the GirlsMPowered program. Utilizing data from 2015 to 2023, including program and race participant datasets alongside Census data, logistic regression, decision trees, and random forests were employed to predict race participant check-ins. Hyperparameter tuning techniques were utilized to prevent overfitting. Additionally, the influence of demographic factors on participant attendance was examined using Kaplan-Meier and Cox Proportional Hazard analyses, revealing significant one-year participation trends. K-Means clustering aided in identifying potential areas for GirlsMPowered program expansion. Overall, these methods contribute to understanding participant demographics and enable accurate predictions of event attendance, informing decisions on program expansion initiatives.

Introduction

The United Neighborhood Organization (UNO) of Chicago is a registered 501(c)(3) nonprofit organization which empowers Hispanic communities through education, advocacy, community development, youth programs, and civic engagement initiatives.

Carrera de los Muertos® (Race of the Dead 5K)

- UNO owns the trademark to the cultural race
- Uses proceeds from the race to fund other UNO programs

First Race in 2007 — First held in Pilsen, IL

GirlsMPowered Program

- Empower young girls in building soft skills socioemotional skills
- Providing exposure to career opportunities not traditional to women

Aged 10-16 — Exposure to web tools — Career Exposure

Business Problem Framing

Carrera De Los Muertos (Race of the Dead 5K)

UNO is looking to expand the program as much as possible given its cultural niche. However, given time constraints for earlier event planning:

- UNO needs to place T-shirt orders and book venue earlier in the year
- Attendance is unpredictable year to year, especially with the previous instances where covid has happened

How big of a venue do they need to rent out to cater for all the participants?

Who are their participants and where are they coming from? Trends over the years?

GirlsMPowered Program

UNO wants to help more girls in need to achieve better lives. However, given that the program is relatively new:

- UNO needs to tell a compelling story to secure funding for program expansion
- UNO needs a concrete direction on where they should expand

How can UNO tell a better story on the communities being served by their programs?

Which other areas can UNO expand to serve?

Analytical Problem Framing

Carrera De Los Muertos (Race of the Dead 5K)

To address these challenges effectively, it's crucial to gain insights into race participants and forecast their check-ins on the event day. With participant data collected well in advance, two possible approaches are:

Approach 1:

Participant earlier cutoff time to collect registration data ahead of time → Utilizing historical participant data, model predicts race day check-ins for each participant. → Understand better, the predicted number of participants on the day of the event, include buffer for last minute registrations → Book and purchase supplies/ venue without having to waste resources

Approach 2:

Analyzing past race data to visualize participant trends and demographics.

GirlsMPowered Program

To address these problems, it's important to show the needs of program participants and find other zip codes that belong to similar clusters as the area of focus. By collecting relevant census data, the two possible approaches are:

Approach 1:

Collect Census data → Clustering zip codes based on program set criteria → Expanding to help more girls in need

Approach 2:

Visualizing poverty levels and racial break down in comparison to national level

Methodology

Data

Race of the Dead 5K
- Shirt information transformed to 3 columns for easier processing
  - Shirt size, shirt material, shirt type
- Imputed missing values using mode for zip code, and KNN for registration method
- Deriving age= [event date]- [date of birth]
- Unique Participant ID: **firstnamelastname_DOByear**

GirlsMPowered Program
- Collecting different attributes per zip code
- Converted to percentages of total population

Exploratory Data Analysis

Race of the Dead 5K
- Visualizing participant demographics and shirt demand

GirlsMPowered Program
- Visualizing the lower west side demographics

Feature Selection

Race of the Dead 5K
- Dropping event specific variables such as donation amounts, event ID, waiver signed, discount name given that model focus is on identifying traits in participant that possibly contribute to event participation check in

GirlsMPowered Program
- Per discussion with UNO, removing aggregate, keeping only variables pertaining to the themes of Hispanic/ African American, Below Poverty Level, and girls aged between 10-19 and highest education attainment level in the household

Modelling

Race of the Dead 5K
- Logistic Regression, Decision Tree and Random Forest to predict for participant check in
- Survival Analysis was conducted to identify key features influencing the risk of participants not showing up

GirlsMPowered Program
- Unsupervised K-Means clustering to group similar zip codes

Model Evaluation

Race of the Dead 5K
- Identifying the best hyperparameters via GridSearchCV or Random SearchCV
- Model performance evaluated based on F-1 Score, Accuracy, ROC-AUC Score

GirlsMPowered Program
- Identifying the ideal number of K based on silhouette method and elbow method
- Cross referencing grouped Zip Codes with information provided on government website on general characteristics such as poverty level

Data

UNO provided past year data for the Race of the Dead 5K program whereas the GirlsMPowered Program was mainly based on Census data of zip codes in Chicago.

Carrera de los Muertos® (Race of the Dead 5K)

40,028 Instances **80 Variables**

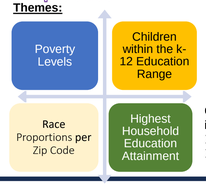
The registration dataset spans from 2015 to 2023, excluding 2021(canceled)

Consisting of information from:
- Financial Summaries
- Final Registration Records
- Price Reports

GirlsMPowered Program

59 Chicago Zip Codes **493 Variables**

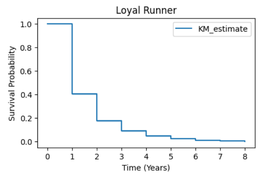
Themes:



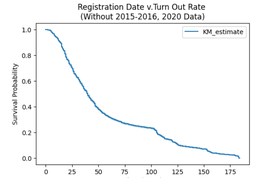
Consisting of information from:
- Program Survey
- Census Data

Model Building & Results

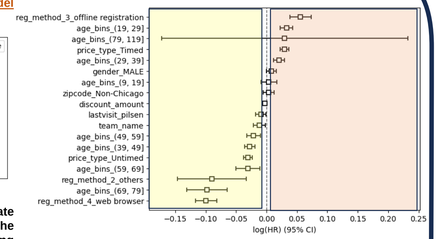
Carrera De Los Muertos (Race of the Dead 5K)- Kaplan Meier and Cox Proportional Hazard Model



60% of the runners are likely to only join for the first year, with 20% of the runners being likely to join for the second year of the race



The further away the registration date from the actual event, the lower the probability of the participant showing up



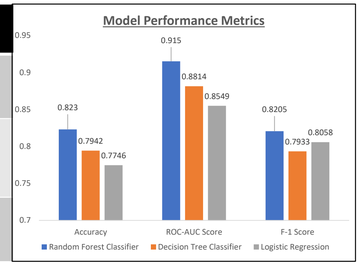
From here, factors that are **above or below the 0 value** influence the participant check in on the day of the event:

- Users tend are more likely to check in if they registered using the web browser (via Active)
- Younger participants (aged 19-29) are more likely to cancel as compared to the other participants
- Males are slightly more likely to cancel on check in as compared to females
- Team runners are more likely to show up for the event as compared to individual runners
- There is significant uncertainty regarding the participation of individuals (aged 79-119) due to a wider 95% confidence interval gap

Carrera De Los Muertos (Race of the Dead 5K)- Participant Check In Prediction Model

Graph below summarizes model performance per metric , Random Forest proves to be the best in predicting participant check in on the day of the event

Model	Best Performing Parameters (Based on GridSearchCV versus RandomSearchCV)
Logistic Regression Best CV score: 0.7764	{C=2, penalty='l1', solver='liblinear'}
Decision Tree Best CV score: 0.7942	{criterion='entropy', max_depth=9, min_samples_leaf=4, min_samples_split=10, splitter='random'}
Random Forest Best CV Score: 0.8534	{Max_depth=15, min_samples_leaf=1, min_samples_split=6, n_estimators=70}



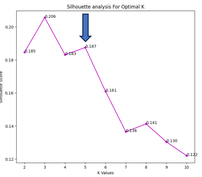
Softwares Used:

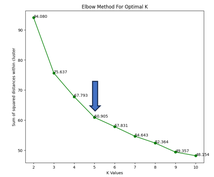
tableau

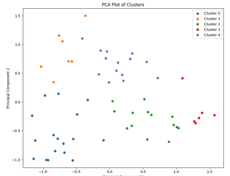
- Scikit Learn
- MatPlotLib
- Scipy
- Imblearn
- Lifelines

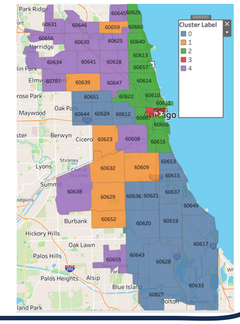
GirlsMPowered Program- KMeans Clustering Results

- K (number of clusters) was determined using the intersection of the silhouette score and elbow method
- Zip codes have been grouped in terms of similarity of the criterias listed in the dataset
  - Zip Codes with similar demographic to 60608 are highlighted in purple on the map to the right (cluster number 4)









Deployment & Life Cycle Management

- Formal documentation of the process was created to ensure replicability of the outputs
- Code and Files for the project shared with sponsor:
  - Merged dataset
  - Tableau TWBX Files
  - Python Code

Conclusion

Carrera De Los Muertos (Race of the Dead 5K)

- Random Forest proves to be the best model to predict for participant Check-In the day of the event with overall highest performance across all model evaluation metrics

GirlsMPowered Program

- Zip Codes for Program Expansion:
  - Garfield Ridge (60638)
  - Mount Greenwood (60655)
  - Logan Square (60647)
  - Edison Park (60631)
  - Roscoe Village (60618)
  - Portage Park (60641)
  - Forest Glen (60646)
  - Elmwood Park (60707)
  - Harwood Heights, Harwood Hts, Norridge (60656)
  - West Ridge (60645)
  - Dunning (60634)
  - Lincoln Square (60625)
  - Jefferson Park (60630)

Acknowledgement

We would like to thank UNO for the opportunity to apply our knowledge in helping communities in need and UIC MSBA Program for providing guidance to complete this project.

To see more on the implementation: