Business Analytics Consultant Project for UMD Alumni Association

Team 4 – Red Turtle

Members： Adarsh Challa, George Puthean, Po-Han Yen (Team Leader), Yi-Wei Sun

Objectives:

- Help the UMD alumni association find ways to attract more first time attendees and major prospects.
- Predict the percentage first time attendees and percentage major prospect based on the characteristics of the event that is going to be held.

Python Libraries used:

numpy, pandas, matplotlib, seaborn, calendar, sklearn

Target value:

Percentage first time attendees (PFTA) & Percentage major prospect (PMP)

Methods used and business analytics processes:

Read the data set using pandas → read_excel

Construct two pandas dataframes for analyzing PFTA (Exclude 2013 data) and PMP using pandas → concat

Exploratory data analysis on numerical data

Explore statistics using pandas → describe

Explore correlations using seaborn → pariplot

Scatter plot between average age and PMP using matplotlib → scatter, title, xlabel, ylabel

Scatter plot between PMP & PFTA using matplotlib → scatter, title, xlabel, ylabel

Exploratory data analysis on time series

Plot Percentage First Time Attendees vs Event Data using seaborn → lineplot

Plot Percentage Major Prospect vs Event Date using seaborn → line plot

Adding column year using pandas → DateTimeIndex

Adding column month using pandas → DateTimeIndex

Adding column weekday using pandas, calendar → DateTimeIndex, day_name

Group by year and plot the mean of PFTA vs Year using pandas and matplotlib → groupby

Group by month and plot the mean of PFTA vs Month using pandas and matplotlib → groupby

Group by weekday and plot the mean of PFTA vs Weekday using pandas and matplotlib → groupby

Group by year and plot the mean of PMP vs Year using pandas and matplotlib → groupby

Group by month and plot the mean of PMP vs Month using pandas and matplotlib → groupby

Group by weekday and plot the mean of PMP vs Weekday using pandas and matplotlib → groupby

Exploratory data analysis on categorical data

Explore statistics of PFTA using pandas → describe

Explore unique values of 'Activity Code', 'Location Code', 'Group Code' using pandas → nunique

Explore statistics of PMP using pandas → describe

Explore unique values of 'Activity Code', 'Location Code', 'Group Code' using pandas → nunique

Analyzing percentage first time attendees

Explore statistics of Mean PFTA by location using pandas → describe

Explore locations with the highest PFTA (Greater than 75 percentile) using pandas and matplotlib → groupby, query, sort_values, index, bar, title, xticks

Explore locations with the lowest PFTA (Lower than 25 percentile) using pandas and matplotlib → groupby, query, sort_values, index, bar, title, xticks

Explore statistics of Mean PFTA by group using pandas → describe

Explore groups with the highest PFTA (Greater than 75 percentile) using pandas and matplotlib → groupby, query, sort_values, index, bar, title

Explore groups with the lowest PFTA (Lower than 25 percentile) using pandas and matplotlib → groupby, query, sort_values, index, bar, title

Analyzing percentage major prospect

Explore statistics of Mean PMP by location using pandas → describe

Explore locations with the highest PMP (Greater than 75 percentile) using pandas and matplotlib → groupby, query, sort_values, index, bar, title, xticks

Explore locations with the lowest PMP (Lower than 25 percentile) using pandas and matplotlib → groupby, query, sort_values, index, bar, title, xticks

Explore statistics of Mean PMP by group using pandas → describe

Explore groups with the highest PMP (Greater than 75 percentile) using pandas and matplotlib → groupby, query, sort_values, index, bar, title

Explore groups with the lowest PMP (Lower than 50 percentile) using pandas and matplotlib → groupby, query, sort_values, index, bar, title, xticks

Further analysis on categorical data

Define a fuction that returns a category based on location code and apply this function to location code column to get location column with 7 unique locations

Then plot out the locations vs PFTA & PMP

Define a fuction that returns a category based on group code and apply this function to group code column to get group column with 4 unique groups

Then plot out the groups vs PFTA & PMP

Define a fuction that returns a category based on activity description and apply this function to activity description column to get activity column with 4 unique activities

Then plot out the activities vs PFTA & PMP

Predictive Model Building

Select relevant variables that are correlated to our target values using pandas indexing.

Split the data into training and testing using sklearn model selection → train_test_split

Convert categorical data into dummy variables using sklearn preprocessing → OneHotEncoder

Combine the converted categorical data with numerical data using pandas → concat

Build and train the models using sklearn linear_model → LinearRegression, fit, predict

Evaluate the models using sklearn metrics → mean_absolute_error

Accomplishments:

Successfully generated insights from data set

1. Discover that PFTA does not highly correlated with PMP

2. Discover the key factors that influences PFTA:

Average Age, year, month, location, group, activity

3. Discover the key factors that influences PMP:

Year, month, weekday, location, activity

4. Discover ways to improve PFTA:

I. Events should be held on May, August, September and October.

II. Preferable locations are international and on campus.

III. Preferable types are social and athletics.

IV. Preferable activity -- SALC.

5. Discover ways to improve PMP:

I. Events should be held on May and September, preferably Thursday.

II. Try hold at South East region and prevent online events.

III. Try to attract alumni with higher age.

IV. Preferable activity -- Dinner. Unpreferable -- SALC, AA.

Build two predictive models with one having better performance

Able to predict PMP quite accurately in the future.

Model works even if the location, group or activity codes we get in the future do not exist in the original data set because we are able to categorize them based on similarities.

Identify ways to further generate insights and improve the models

Group activity codes into smaller groups based on similarities and domain knowledge.

Try different models to find the one with higher prediction accuracy.

Construct pipelines to simplify data cleansing and model training processes.

Testing and evaluating:

Sequence based on presentation slides

Check if the methods used are correct.

Check if the correlations between numerical variables are correct.

Check if the label and titles match each plot.

Check if the ticks on the labels make sense.

Check if the stated insights are logical enough.

Check if the plots can deliver the intended massages clearly.

Check if the methods used to generate new columns: year, month, weekday, location, group, activity make sense.

Check if the codes contain any errors.

Check if the recommendations are consistant with the insights.

Check if the data preparing, model building and evaluating proccesses are correct.

If we get more data in the future, used them to check the model predictions are close to the MAE.

Dictionary used for further analysis on categorical data:

Group

Athletics: PA9, PAB, PAG, PAY, PAZ,

Social: P99, PS3, PS9, PSA, PSB, PSC, PSK, PSL, PSS, PST, PSX, PSY, PSZ,

PRODEV: PC10, PC11, PC4, PC9, PCC, PCS, PCW, PCZ,

Others:

ADVOCACY: PD9,

Campaign: PG9,

Stewardship: PH9, PHM,

CULTIVATION: PI9, PIZ,

MEMBERSHIP: PM9, PMM,

SERVICE: PO9, POB, POC, POO, PSO

AFFINITY: PQB, PQK, PQL, PQX,

CULTURAL: PU9, PUB, PUL, PUU, PUZ,

PVW

Location

DMV: PDAN, PDBA, PDBR, PDCP, PDDC, PDES, PDHC, PDMC, PDNA, PDNV, PDPG, PDSO, PDWE

ON CAMPUS: PDON

INTERNATIONAL: PISK, PIUK, P9NA

NORTHEAST: PNBO, PNCH, PNCI, PNDE, PNNA, PNNJ, PNNY, PNPH,

Online: POBR, POWE,

SOUTHEAST: PSAT, PSAU, PSDF, PSHO, PSJA, PSMF, PSNA, PSRT, PSTA,

WEST COAST: PWDE, PWDF, PWLA, PWNA, PWOC, PWSD, PWSE, PWSF, PWTU

Activity

CP AAE

CP AA

CP SALC

CP Dinner