Case Study Summary Session

MIT-ADSP

Topics

Foundations of Data Science

Inferential Statistics

Topics

Inferential Statistics - Medicone Dose Testing Case Study

Objective and dataset

- Medicone, a pharmaceutical company, has manufactured the sixth batch of COVID-19 vaccine doses.
- They want to understand the quality of the vaccine with the help of some voluntary trials.
- Find the quality and effectiveness of the vaccine so that the company can plan for the next batch of doses.

Approach

- Read the description of the problem and find some important properties from the results of previous studies.
- Plot the probability distribution for doses and find the results from PMF and CDF.
- Load the dataset and use some visualization to find the data distribution type.

Key Findings

- The probability that the vaccines will not do a satisfactory job is very low.
- The probability that exactly 3 doses will not do a satisfactory job is 1.25%, and the probability that at most 3 doses will not do a satisfactory job is 1.73%.
- The probability that at least 30 doses will not do a satisfactory job out of 200 doses is ~0.3%.
- 95% of the time, the mean effective time of doses will be between 12.09 hours and 14.79 hours.

Foundations of Data Science

Hypothesis Testing

Topics

Hypothesis Testing - Mobile Internet Case Study

Objective and dataset

- ExperienceMyServices reported that a typical American spends an average of 144 minutes (2.4 hours) per day accessing the Internet via a mobile device with a standard deviation of 110 minutes. Test the validity of this statement.
- Dataset consists of 30 samples and only the value of time spent per day accessing the Internet .
- Find out if there is enough statistical evidence to conclude that the population average time spent per day accessing the Internet via mobile device is different from 144 minutes?

Approach

- Define null and alternate hypotheses and decide the significance level.
- Identify the test statistic - z-statistic or t-statistic.
- Calculate the p-value using the test statistic.
- Decide whether to reject the null hypothesis or not based on the test statistic.

Key Findings

- At 5% significance level, we do not have enough statistical evidence to prove that the average time spent on the Internet is not equal to 144 minutes.

Data Analysis and Visualization

PCA and t-SNE

Topics

Air Pollution Case Study

Objective and dataset	• Reduce the number of features using dimensionality reduction techniques like PCA and t-SNE • Apply dimensionality reduction techniques on AirPollution dataset and interpret/visualize the results • Air pollution dataset containing information on major pollutants and meteorological levels of a city
Approach	• Load the air pollution dataset and drop serial number and date columns • Check the distribution and outliers for numerical columns in the data • Calculate the correlation coefficients between different variables • Check and impute missing values and do the scaling of the data • Reduce dimensions using PCA and t-SNE and visualize the data in 2 dimensions
Key Findings	• In the air pollution dataset, reduced the number of features by ~82% (from 28 to 5) using PCA with 30% loss in variance • Visualized the data in lower dimensions using t-SNE • The data forms 4 groups described below: ○ Group 1 represents hot and humid areas ○ Group 2 represents developing urban areas ○ Group 3 represents the developed urban areas ○ Group 4 represents the industrial areas

Data Analysis and Visualization

Network Analysis

Topics

Caviar Investigation Phases Case Study

Objective and dataset

- A time-varying criminal network that is repeatedly disturbed by police forces from 1994 to 1996 in eleven phases
- The network consists of 110 (numbered) players. Players 1-82 are the traffickers. Players 83-110 are the non-traffickers (financial investors; accountants; owners of various importation businesses, etc.)
- Understand, create and visualise the data in phases and figure out important nodes across phases

Approach

- Read the data and understand the structure of data
- Put the data into a graph and visualize the graph
- Identify the important nodes from the visualization
- Calculate the centrality measures (Degree, Eigen, Betweenness, Closeness) and quantify the importance
- Understand the variation of node importance across phases

Key Findings

- We carried out the analysis on the network and figured out techniques to read adjacency matrices into graphs
- We later visualised the graphs, created centrality measures and identified important nodes - N1, N3, N12
- We studied and plotted the variation in the centrality of the important nodes across phases in a bid to understand the effect of disruption of the network

Data Analysis and Visualization

Unsupervised Learning

Topics

Country Clustering Case Study

Great Learning
POWER AHEAD

Objective and dataset	• Identify cluster of countries that are more similar to each other in terms of certain socio-economic factors • Country dataset contains various socio-economic attributes for countries around the world • We will not do clustering on the gdp and would rather try to understand the variation of other factors with GDP across the groups that we formed
Approach	• Load the country dataset and perform basic univariate analysis • Check correlation among numerical variables and scale the data • Choose best K using elbow method and Silhouette score and create cluster profiles using the K-Means clustering • Apply K-Medoids clustering and compare the cluster profiles with K-Means clustering profiles • Apply Gaussian Mixture clustering and compare the cluster profiles with K-Medoids clustering profiles • Choose the number of clusters using the Dendrogram and apply hierarchical clustering • Apply DBSCAN clustering
Key Findings	• No clear 'elbow' in the elbow plot but the Silhouette score is highest for K=3. Chosen K=3 for K-Means • Cluster 2 has only 3 observations. It consists of outlier countries with highest imports and exports as percentage of GDP • Cluster 1 shows traits of underdeveloped and developing countries and Cluster 3 shows traits of developing and developed countries • Using K-Medoids, cluster 2 represents underdeveloped to developing countries, cluster 1 represents developing countries and cluster 3 represents developed countries • The count of observations in each clusters from K-Medoids is more evenly distributed as compared to clusters K-Means • Unlike K-Means, the clusters from K-Medoids for developed countries is much bigger but still retains the overall characteristics of developed countries • In GMM, clusters looks very similar to the clusters from K-Medoids with one cluster of 'rich' countries, one of 'poor' and one of 'all others'. 0, 1, and 2 represents underdeveloped, developed, and underdeveloped and developed countries, resp. • It is hard to distinguish clusters using hierarchical clustering. Therefore, we will not deep dive into the cluster profiles. • In DBSCAN, we got 5 clusters using epsilon equal to 1. Three out of 5 clusters (0,1,& 2) seems to be way more compact across all attributes. We can explore it more to understand which type of countries it consists. • Choice of algorithm here will depend on the context and use case. But purely based on foundations of 'what good clustering looks like', one can propose K-Medoids as it has more distinct extreme clusters of developing and underdeveloped countries.

Machine Learning

Regression and Model Evaluation

Topics

BigMart Sales Prediction Case Study

Objective and dataset

- Data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities
- Build a predictive model and find out the sales of each product at a particular store
- Provide recommendations to the BigMart sales team to understand the properties of products and stores which play a key role in increasing sales

Approach

- Load the dataset and drop ID variables - Item_Identifier, Outlet_Identifier
- Perform univariate and bivariate analysis. Check correlation among numerical variables
- Fix the data issues in the column Item_Fat_Content, impute missing values and create new feature Outlet_Age
- Prepare data for modeling and scale the training and testing data
- Build the model and select only the relevant features based on p-value (p-value<0.05)
- Check for 5 assumptions of the linear regression model
- Conclusion and Recommendations

Key Findings

- The majority of Outlet_Size is Medium, majority of Outlet_Location_Type is Tier 3, and majority of Outlet_Type is Supermarket Type 1
- The average sales are almost constant every year except 1998 where the average sales plummeted
- Age of stores does not impact the sales as different age of stores have similar distribution approximately
- After removing multicollinearity, applying log transformation on the target variable, and checking all the assumptions, the final model is giving the R-Square of 0.675
- The R-Squared and MSE on the cross validation is almost similar to the R-Squared on the training dataset

Conclusions and Recommendations

- Equation of the model implies one unit change in the variable Item_MRP, the outcome variable increases by 1.9623 units.
- On average, the log sales of stores with outlet size small is 0.5812 less than the log sales of outlet size high
- On average, the log sales of store type Supermarket 3 is more than the log sales of other types of stores.
- The management can focus on maintaining or improving the sales in large stores of supermarket type 3. And for the remaining ones we may want to make strategies to improve the sales e.g. better training for store staffs, providing more visibility of high MRP item, etc.

Machine Learning

Classification

Topics

Employee Attrition Case Study

Objective and dataset	• McCurr Healthcare Consultancy is an MNC that has thousands of employees spread out across the globe • The Head of People Operations wants to bring down the cost of retaining employees • Identify the factors that drive attrition and build a model to predict if an employee will attrite or not • The data contains employee information like demographic details, work-related metrics and attrition flag
Approach	• Load the dataset and drop unnecessary columns • Perform univariate and bivariate analysis. Check correlation among numerical variables • Prepare data for modeling and scale the training and testing data • Build the model using different algorithms - Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression, and KNN. • Interpret the results and print the classification metrics for the training and testing data
Key Findings	• Working overtime is the most important driver of attrition • Attrition rate is high for sales and marketing departments • The organization has a lower percentage salary hike and promotions are given less frequently • Approximately 40% employees have given a poor rating on job satisfaction and environment satisfaction • Lower job involvement leads to a higher likelihood of attrition • Young and relatively new/inexperienced employees tend to show a higher attrition rate
Conclusions and Recommendations	• The hyperparameter tuned KNN classifier is overfitting but gives the highest recall on the training and the testing data • The organization should manage their work more efficiently so that employees don't have to work overtime and can manage to have a work-life balance, or failing this, the company could provide some additional incentives to employees who are working overtime • The organization could look into their incentive schemes and try to come up with better ideas to retain employees from sales and marketing departments • The company might be able to focus on giving promotions more frequently or they could increase the annual appraisal hike. Also, a more proactive, hands-on approach may be required from the managers in the organization to avoid low job involvement

Appendix

Topics

UBER Case Study

Objective and dataset	• Find out the different factors that influence pickups and the reason for such influence • To identify some key insights that Uber management can take reference from to capitalize on fluctuating demand. • The Uber dataset contains various information about weather, location, and no. of pickups.
Approach	• Load the dataset, fill in the missing values and drop unnecessary columns. • Perform univariate and multivariate analysis. Explore categorical variables and check correlation among numerical variables. • Draw meaningful conclusions from the different plots.
Key Findings	• Uber cabs are most popular in the Manhattan area of New York. • Contrary to intuition, weather conditions do not have much impact on the number of Uber pickups. • The demand for Uber increased steadily over the months (Jan to June). • The rate of pickups is higher on the weekends as compared to weekdays. • People use Uber for regular office commutes. The demand steadily increases from 6 AM to 10 AM, then declines a little and starts picking up till midnight. The demand peaks at 7-8 PM. • New Yorkers trust Uber taxi services when they step out to enjoy their evenings.
Conclusions and Recommendations	• Manhattan is the most mature market for Uber. Brooklyn, Queens, and Bronx show potential. • There has been a gradual increase in Uber rides over the last few months and we need to keep up the momentum. • Ridership is high at peak office commute hours on weekdays and during late evenings on Saturdays. Cab availability must be ensured during these times. • The demand for cabs is highest on Saturday nights. Cab availability must be ensured during this time of the week. • Procure data for fleet size availability to get a better understanding of the demand-supply status and build a machine learning model to accurately predict pickups per hour, to optimize the cab fleet in respective areas. • Procure more data on price and build a model that can predict optimal pricing.