Case Study Summary

MIT ADSP

Topics

1. Data Analysis and Visualization
    1.1. PCA and TSNE
    1.2. Network Analysis
    1.3. Unsupervised Learning
2. Machine Learning
    2.1. Regression and Model Evaluation
    2.2. Classification
3. Practical Data Science
    3.1. Decision Tree and Random Forest
    3.2. Time Series

Data Analysis and Visualization

PCA and t-SNE

Topics

Education and Air Pollution Case Study

Great Learning
POWER AHEAD

Objective and dataset

- Reduce the number of features using dimensionality reduction techniques like PCA and t-SNE
- Apply techniques on two datasets and interpret/visualize the results
- Education dataset containing information on educational institutes in USA
- Air pollution dataset containing information on major pollutants and meteorological levels of a city

Approach

- Load the education dataset and perform basic univariate analysis
- Fixed percentage values greater than 100 for two columns
- Check correlation among numerical variables and scale the data
- Reduce dimensions using PCA and t-SNE and visualize the data in 2 dimensions
- Load the air pollution dataset and drop serial number and date columns
- Check and impute missing values
- Reduce dimensions using PCA and t-SNE and visualize the data in 2 dimensions

Key Findings

- In education dataset, reduced the number of features by ~76% (from 17 to 4) using PCA with 30% loss in variance
- Represented each principal component as a linear combination of original features
- PC1 captures attributes that define premier colleges with high quality of students entering them and higher accomplishing faculty that is teaching there. They also seems to take rich students from all over the country.
- PC2 captures attributes that generally define non-premier colleges that are comparatively easier to get admissions into
- PC3 is related to financial aspects and low values of student faculty ratios
- PC4 captures attributes that define colleges with lack the highly educated faculty but it is comparatively easier to graduate from there
- No meaningful pattern observed using t-SNE. The data is clustered together with some outliers
- In air pollution dataset, reduced the number of features by ~80% (from 25 to 5) using PCA with 30% loss in variance
- Visualized the data in dimensions using t-SNE
- The data forms 4 groups where Group 1 represents hot and humid areas, Group 2 represents developing urban areas, Group 3 represents the developed urban areas, and Group 4 represents the industrial areas

Data Analysis and Visualization

Network Analysis

Topics

Caviar Investigation Phases Case Study

Great Learning
POWER AHEAD

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

Great Learning
POWER AHEAD

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

Great Learning
POWER AHEAD

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 thet 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

Practical Data Science

Decision Trees and Random Forest

Topics

Employee Attrition Case Study

Great Learning
POWER AHEAD

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
- Encode the categorical variables and split the data into training and testing data in 70:30 ratio
- Build the model decision tree and random forest models and analyze the performance on the training and testing data
- Visualize the decision tree and check the feature importance of both the models
- Try to improve the model performance using hyperparameter tuning with GridSearchCV

Key Findings

- The Decision Tree model with default parameters is overfitting the training data
- Tuning the decision tree model reduced overfitting but recall on the test data decreased significantly
- The Random Forest classifier is overfitting the data as well. Recall on the test data is about 79%
- The tuned random forest model is also comparatively overfitting the training dataset, but it shows a very good performance on the test dataset. The recall for the tuned model has improved from 79% to 83% with a small decrease in precision
- The feature importance plot for tuned random forest model suggests that OverTime, MonthlyIncome, Age, TotalWorkingYears, and DailyRate are the most important features

Conclusions and Recommendations

- The tuned random forest model is the best model with nearly ~83% recall on the test data. The company can use this model to know beforehand which employee is going to attrite and act accordingly
- 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 company should make sure that all its employees are compensated at least based on industry standards
- The company should also keep track of the hourly rate or the daily rate, so that when the employees need to stay overtime for extra work they are well compensated for the same

Practical Data Science

Time Series

Topics

Bitcoin Price Prediction Case Study

Great Learning
POWER AHEAD

Objective and dataset

- Bitcoin is a decentralized digital currency that uses encryption schemes, and other mechanisms to verify transactions and ensure security
- An unusual feature of Bitcoin price is the large fluctuation in contrast to traditional financial assets
- The dataset consists of monthly average closing prices of Bitcoin from Dec 2011 to March 2021
- Build a time series model using the AR, MA, ARMA & ARIMA models to predict the monthly average closing price of Bitcoin

Approach

- Load the dataset and visualize the time series
- Split the dataset in train and test data. Keep last 12 months as the test data and test the stationarity of the training data
- Perform log transformation and differencing to make the data stationary and visualize the demposed data
- Find values of p and q using acf and pacf plots
- Build the model using different algorithms - AR, MA, ARMA, and ARIMA
- Inverse transform the predictions
- Forecast the series for last 12 months (from April 2020 to March 2021) and visualize the same

Key Findings

- The time series has an upward trend with some seasonality. The value of Bitcoin has also increased tremendously in the year 2021 alone
- The seasonality of Bitcoin shows that its price spikes from December to January and then drops constantly till May
- The acf and pacf plots suggests that p=q=7
- After building all four models, the ARMA model is giving the least RMSE and the second lowest AIC value among all the models
- After forecasting, we observed that most of the predicted values on the training data are close to the actual values except for the spike in the prices in the year 2018 and at the end of 2019
- On the test data, the model is not performing well. the test predictions are not able to identify the volatile variations in the prices over the last 12 months.

Conclusions and Recommendations

- Contrary to our observation, the RMSE is higher on the training data in comparison to the testing data. This might be due to the low number of observations in the testing data
- This might be because there have been many fluctuations in the Bitcoin prices over the last 12 months and our model hasn't learnt this from the pattern of previous years and hence the predicted values are not at all close to the actual values
- Also, our model might not be complex enough to capture these fluctuations - Spikes and Dips. We can further try to build more complex time series models like SARIMA, SARIMAX, etc. while considering more factors like trend, seasonality, etc.