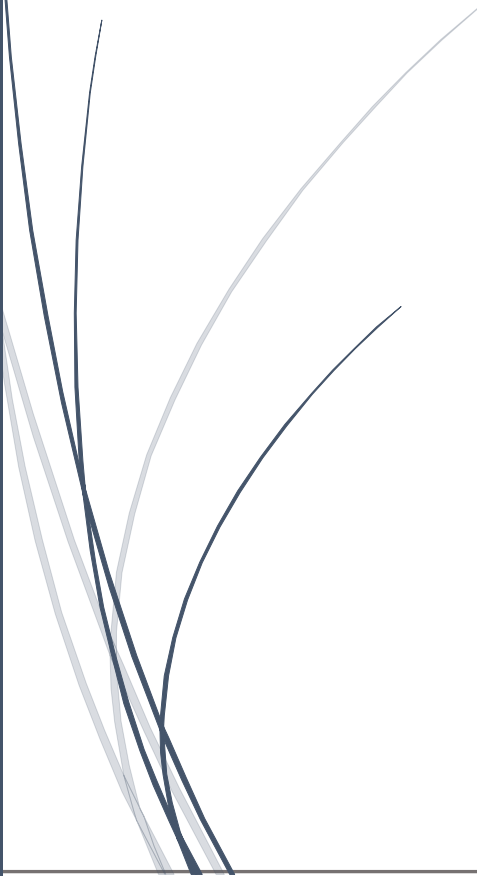


A dark blue vertical bar runs down the left side of the page. A blue arrow points to the right from this bar, containing the date.

3/23/2020

Capstone (DATA 2206)

Module Two (Part II) – Exploratory Data
Analysis (EDA)

Several thin, curved lines in shades of blue and grey originate from the bottom left corner and sweep upwards and to the right.

Kartik Sojitra
100723768

Table of Contents

Problem statement	2
Analytics Rationale Statement.....	2
EDA Action Plan	2
Identify and Justify Output Variable Class Structure	2
Action Plan for Exploratory Data Analysis (EDA).....	2
EDA	3
Data Exploration	3
Data Cleaning	4
Key Insights	4
Analytical Scorecard	7
References	8

Problem statement

Starseed Medical Inc. determine products as the “best” quality products based on the best cannabinoid quality starting from green star 1 to blue star 3 potency CBD cannabis strains. These products will be used for medical so in the classification model, we will consider Recall and f1-score to resolve this problem.

Analytics Rationale Statement

- 1) The cannabinoid quality is measured by the classification of the quality of the plant.
- 2) To achieve the objective, we need to find how can plant provide the best cannabinoid potency to produce the best quality of cannabis-based medical products.
- 3) Starseed Medical Inc. should register the clients to increase the sale of cannabis-based medical products.
- 4) We need more than 95% of Recall value and f1-score as we are using the best quality cannabis for medical use.

EDA Action Plan

Identify and Justify Output Variable Class Structure

To resolve the problem, we are using a binomial classification model to distinguish the best potency from low, average, good and best, and used transformation to determine the best cannabinoid quality output by algorithms.

The output variable for this project will be binomial as our output will be whether we are getting the best quality product of cannabis or not based on the “best” cannabinoid quality starting from green star 1 to blue star 3 potency CBD cannabis strains.

We are choosing binomial structure as we are going to identify only “best” quality products from low, average and good and counting the outcomes as either 1 and 0: counting the best as 1 and others (low, average and good) as 0. All the variables given in the dataset considered as independent variables and utility will be the dependent variable.

Action Plan for Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. EDA focuses on checking assumptions required for model fitting and hypothesis testing and handling missing values and making transformations of variables as needed.

Steps of EDA in Cannabis dataset:

- 1) Importing the required libraries
- 2) Loading the data into the data frame
- 3) Show key statistics (.describe())
- 4) Overview of data characteristics (.info())
- 5) Box plot feature set comparison
- 6) Summary of N/A values (.isnull().sum())
- 7) Drop missing observations and reset index (.dropna())
- 8) Combine classes
- 9) Replace best with '1' and other with '0'
- 10) Handling missing values for future
- 11) Visualization of Correlations (Heat map)
- 12) Density Plot and Histogram of Output Variable
- 13) q-q plot
- 14) Normality test - Shapiro-Wilk Test
- 15) Scatterplot for 'Brnch_Fm' and 'Utility'
- 16) Histogram and Pie chart for data distribution

The Dataset contains raw data originated from the Starseed Medical Inc. and it represents around 100% of testing data from Dec 1st to 10th, 2019. We are going to analyze cannabis dataset to find the key insights, to analyze dependent and independent variables for various types of data, to detect the outliers, to make assumptions and constraints of dataset, handling missing data using the various method like box plot, scatter plot, correlation matrix and statistical imputation method.

EDA

Data Exploration

- No of feature and count of rows.
- Type of values either numeric or string and selecting appropriate datatype for each column.
- Our dataset consists of 1472 rows with 10 features
- We see the distribution of the data in each column -mean, median, mode, max & min value, counts etc.
- Identify independent and dependent variable.
- Box plot will help us find errors/anomalies and outliers in the data which can then be dropped.

Data Cleaning

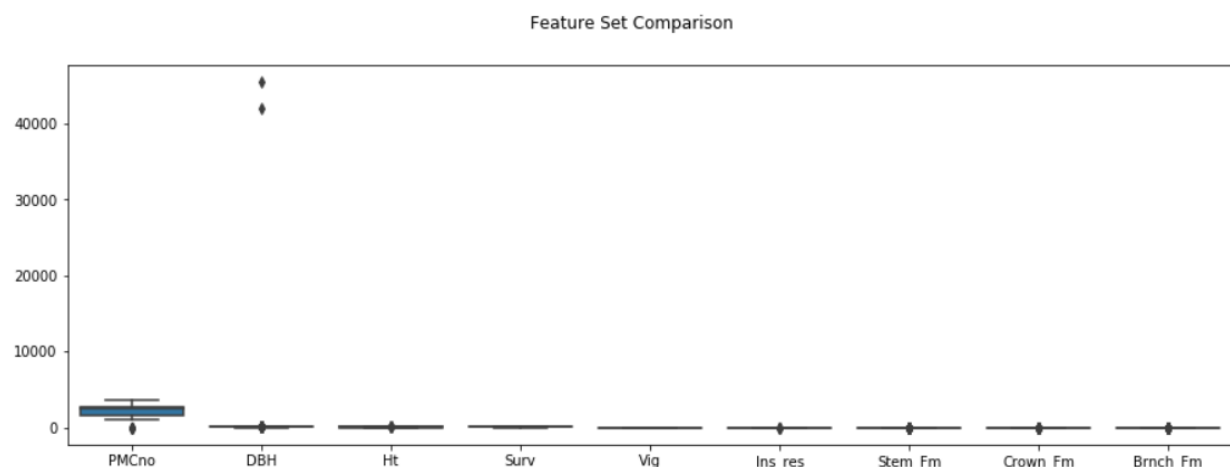
- Summary of N/A values using `.isnull().sum()` in cannabis dataset
- Drop missing observations (Nan) and reset index using `.dropna()`
- Combine classes
- Replace the 'best' as 1 and others (low, average and good) as 0.
- Handling missing values for future
- Split into test and train data.

Our objective is to determine the best quality of products from cannabinoid potency grades from the plant. Data reveals that the cannabinoid quality of the plant is measured by the size, type, and species of the plant. Based on prescriptive analysis it concludes that no need to add more variables to design a better predictive model and achieve high Recall and f1-score. It is very important to know which and how data is used to build a better model and successful data analysis. It also specifies that what makes a good model, how can we validate and optimized the model and which hyperparameter we used to improve the accuracy of the model. Moreover, we can use visualization to understand the correlation between variables.

We are going to use a classification predictive model to map from the independent variable to the dependent variable. We will use the classification algorithm to train the dataset and cluster it. To deal with probability and accuracy we are going to use a logistic regression model.

Key Insights

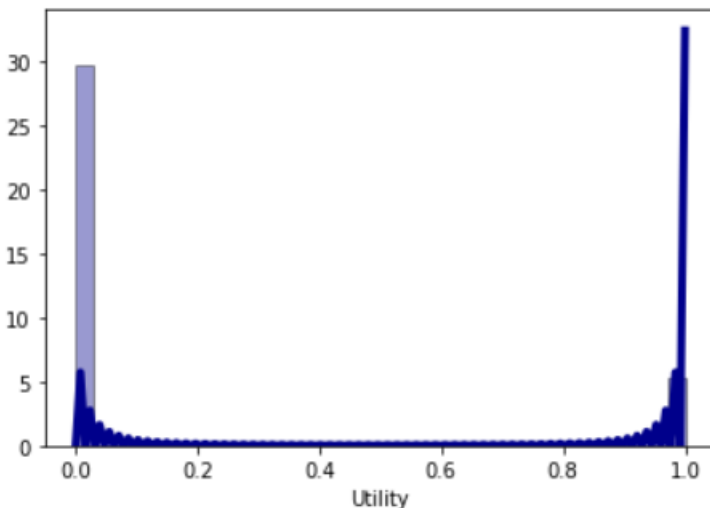
Box plot



This graph helps us understand visually what could be the best range of value for Crown_Fm in case of best strain. From the graph we can see that the value of independent variable lies

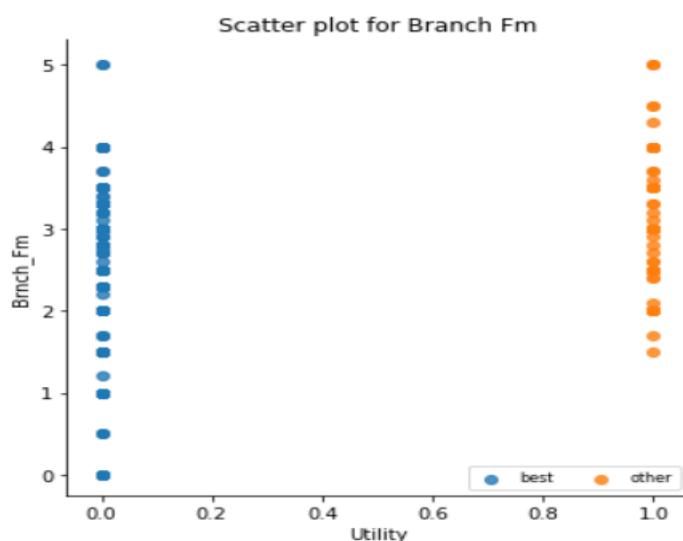
between 0 and 5, and only DBH has few outliers. We are not going to eliminate any column for feature selection as it causes bias during learning and validation. Based on prescriptive analysis it concludes that no need to add more variables to design a better predictive model and achieve high Recall and f1-score.

Density Plot



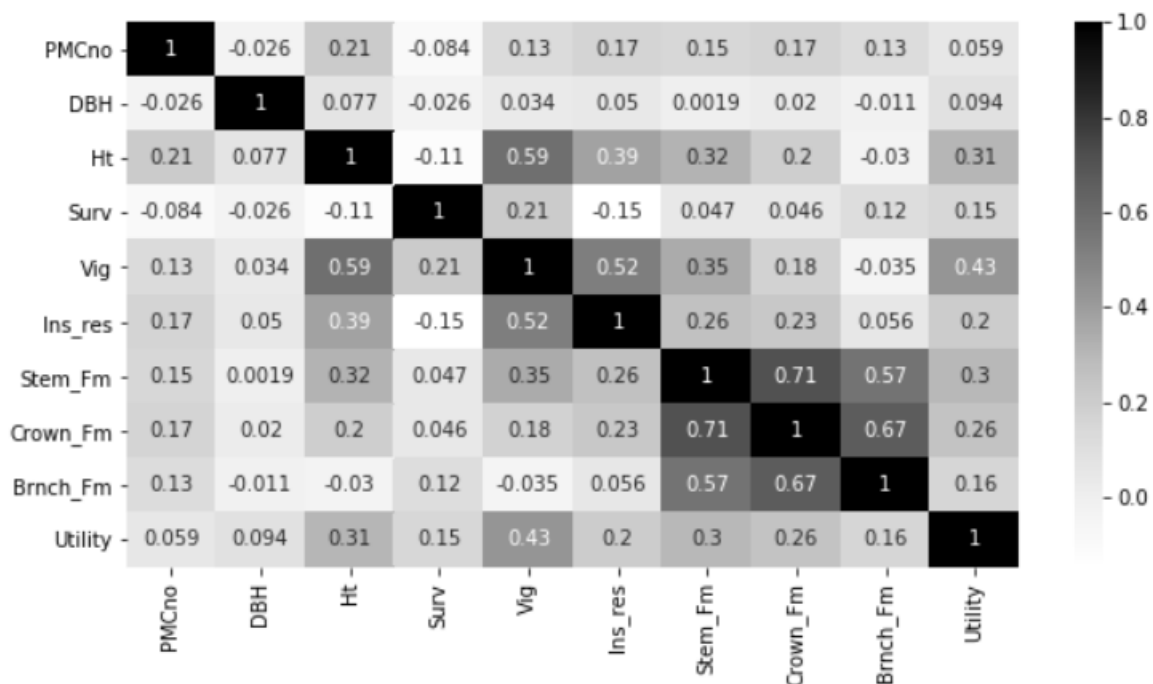
The curve exhibits that there is left and right skewed positive distribution because it has a long tail in the positive direction with few outliers which are shown in the box plot and the mean is to the left of the peak. For those factors whose mean and standard deviation is high they influence the Utility to affect most as compared to a variable whose gap is little.

Scatter plot



Data reveals that the cannabinoid quality of the plant is measured by the size, type, and species of the plant. Scatterplot help us understand relationship between two variables. From the above graph, we can clearly visualize and analyze value of Branch_fm for best quality strain between 2.5 to 3.5. This helps us understand visually what could be the best range of value for Branch_Fm in case of best strain.

Heat Map



Heat Map shows the correlation between the independent variables. It also specifies that what makes a good model, how can we validate and optimized the model and which hyperparameter we used to improve the accuracy of the model. Moreover, we can use visualization to understand the correlation between variables. From the correlation matrix graph, we can see that Vig has strong correlation with Utility by 43% whereas with Ht its correlated by 31%. Additionally, PMCno and DBH has not good correlation with Utility, 5.9% and 9.4% respectively.

Normality test - Shapiro-Wilk Test

Hypothesis statement:

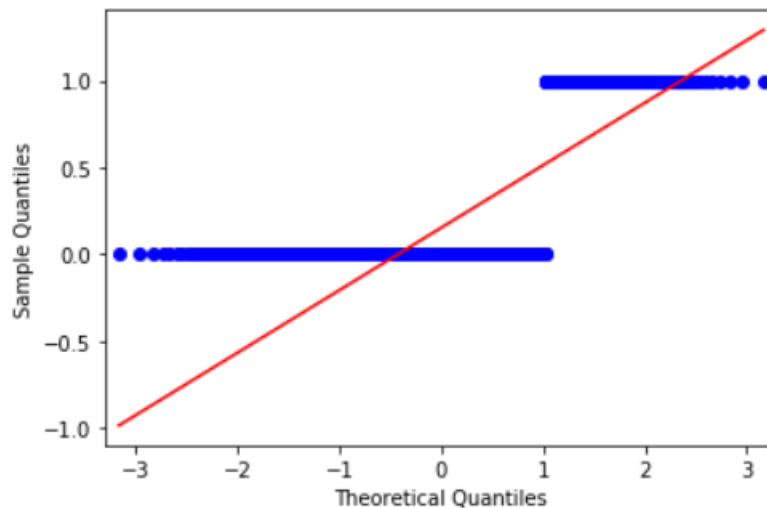
$p \leq \alpha(0.05)$: Reject H_0 , Not normal

$p > \alpha(0.05)$: Fail to reject H_0 , Normal

- After conducting the Shapiro-Wilk normality test we got the p-value 0.00 is less than the alpha (0.05), so we can fail to reject H_0 states that the predictors are not related to the response variable.

- It also says that the sample does not look Gaussian and there is not a normal distribution in the model.

Q-Q Plot



The QQ-plot determines that values of Utility do not adapt very well to the normal distribution. The variation between Utility values and the normal distribution shows to be most noted in the lower left-hand angle of the chart, which compares to the left end of the normal distribution. The inconsistency is remarkable in the uppermost right and left-hand edge of the diagram, which resembles abnormal distribution.

Analytical Scorecard

In order of our model to classify and predict best quality strain composition and plant size of cannabis, which will be used to manufacture high potency medicinal cannabis products we need to consider below key factors.

- **Perspective:**

The perspective of our company is to Increase revenue from bringing premium medical products by Introducing advertisements and market about premium medical products through social media and banners and Reduce costs by growing only the best quality strains through insights gained through data modeling. Also, increase the production of best quality strains starting from green star 1 to blue star 3 potency CBD cannabis strains.

- **Business drivers:**

The data collected should be of high-quality data which captures all the necessary field in order to do the analysis. Any key information missing will largely affect the output. Better decisions come from more reliable and high-quality data about products, services and operations. Data collected and maintained by the company should be both aligned and prepared so that through insights from data several business opportunities.

- **Metrics:**

In the case of modeling our key metric for model evaluation is the confusion matrix. This metric will describe how good our model is performing and is able to come up with the right chemical compound and plant size will the data available.

Key Performance Indicator: KPI is most useful to predict the best quality strain composition and plant size of cannabis strain which can be used to manufacture high potency medicinal cannabis products. Volume of data capturing key features to do the analysis which is able to solve the business problem.

Target Audience: Our target audience will be Vice president of Starseed Medical Inc. This means that insights gained from data modelling must be explained in laymen terms. The scorecard should be concise and non-technical. It should contain KPI, basic measures considered key to monitoring trends in best quality cannabis. Approach taken to come up with result and evaluation metrics to substantiate the findings.

References

Sam Plati_database overview- capstone (Data 2206)

Sam Plati_capstone (Data 2206)_module two (part - 1)_EDA action plan_week 2