DURHAM COLLEGE
SUCCESS MATTERS

4/5/2020

Capstone (DATA 2206)

Module Three – Data Analytics Report

Kartik Sojitra
100723768

Table of Contents

Overview

The purpose of this report is to present the research conducted on cannabis by Starseed Medical Inc, which focusing on bringing premium medical products to the clients of the project which is Starseed Medical Inc decision-makers who will determine products as the “best” quality products based on key findings of the analysis.

The document illustrates the details of Starseed Medical Inc. which is producing the most high-grade quality cannabis to provide premium medical products to their clients. To accomplish the objective Starseed design system that classifies the best variety of cannabis-based on the high cannabinoid quality starting from green star 1 to blue star 3 potency CBD cannabis strains and classifying high potency THC medical cannabis oil. They are examining how do they classify the most useful CBD cannabis strain and the premium products based on cannabinoid potency grade. To resolve the problem, they are using a 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 report is to show key findings from data analysis as well as the sequence of steps implemented in order to address the problem.

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.

Key Questions

The below questions help to understand stakeholder's importance.

Question	Reason it is being to answer
1) How do they define premium product by considering the green star 1 through blue star 3 high potency CBD cannabis strains for the clients and for the decision makers?	The cannabinoid quality is measured by the classification of the quality of the plant. 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.
2) Is an accuracy of 75% or higher a realistic goal considering the data quality and constraints?	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 we need to replace NAs with mean to design a better predictive model and achieve high Recall.
3) Which factors are more important to predict an outcome?	Our objective is to determine the best quality of products from cannabinoid potency grades from the plant. Depending on what would be these factors it can be used clients to increase the sale of cannabis-based medical products.

Dataset Summary

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. The dataset contains the independent variable as an actual measurement parameter for the cannabis plant and a dependent variable that classify the different CBD cannabis strains (none, low, average, good, best).

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.

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 we need to add more variables to design a better predictive model and achieve high Recall and f1-score.

The IT and research department of Starseed Medical Inc. has access of all the data during data collection, data processing and data analysis.

Model Analysis

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) Overview of dataset characteristics (.info())
4) Summery of N/A values (.isnull().sum())
5) Replace NA with Mean, inplace = True mean
6) Combine classes
7) Drop Column
8) Replace best with '1' and other with '0'
9) Handling missing values for future
10) Tukey Method
11) Boxplot of features
12) Class balance – SMOTE
13) Show key statistics
14) Boxplot feature set comparison
15) Plot learning curve
16) Model analysis
17) Optimization using Gridsearch function
18) Run optimized logistic regression model
19) Plot ROC curve

Method

1) Scorecard was set for measuring performance of the project; it includes the following measures:

Measure	Target
Accuracy of the best model built	70% or higher
Precision	80%
Recall	75%

Accuracy is the number of correct predictions made by the model over all kinds predictions made.

Precision of a classification model is its ability to identify only the relevant data points. It is a positive predictive value or exactness. It shows how many out of all positive classes were predicted correctly.

Recall of a model measures completeness. It is how many out of ALL classes (positive and negative) were predicted correctly.

2) Descriptive statistical tools such as data visualization and calculating basic statistics of the data set were performed as a pre-modelling procedure in order to understand data given.

3) For the purpose of this research two statistical methods where chosen: **Logistic regression and Decision tree.**

4) For every model performance metrics were calculated in order to see how effective a model predicts outcome, what is predictive power of every one of them. Performance metrics were used to compare models built and chose the one with the highest predictive power.

5) Finally, possible business outcomes, improvements to the research were stated.

6) After data cleaning, transformation and de-coding the dataset contains 880 observations and 9 columns. 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.

First Algorithm: Logistic regression model as a base model

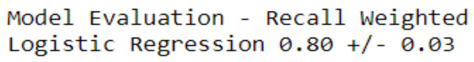
Model Choice

Models to be build were chosen according to the type of dependent factor. I this case, it has only two possible results best and others and we have close dataset, so we have chosen logistic regression model as our base model.

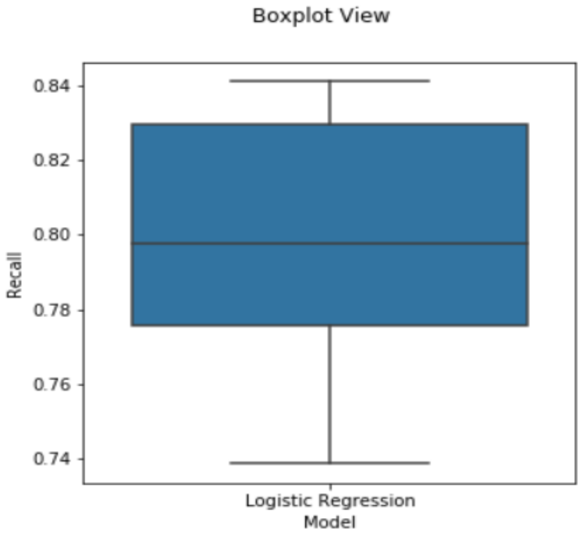
Logistic Regression Model

Logistic Regression is a Machine Learning classification algorithm that is applied to predict the probability of a categorical dependent variable.

In logistic regression, the dependent variable is a binary variable that contains data coded as best (1) and other (0). In other words, the logistic regression model predicts P(Y=1) as a function of X.

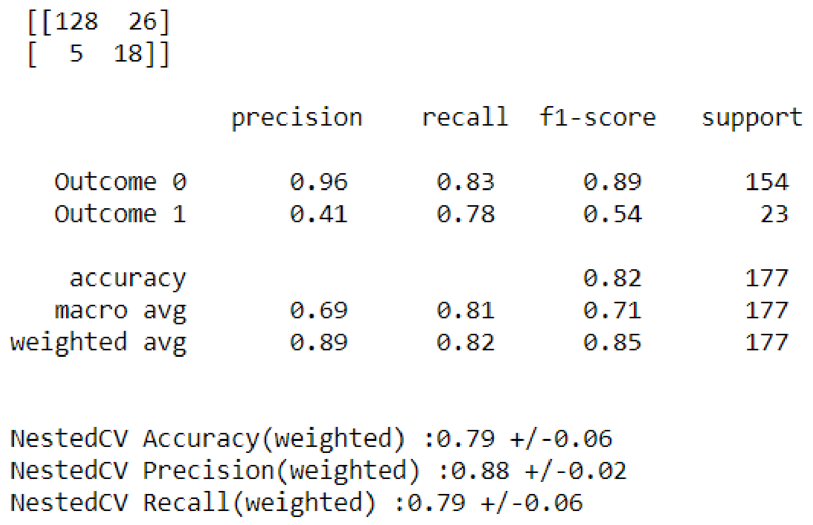


```
Model Evaluation - Recall Weighted
Logistic Regression 0.80 +/- 0.03
```



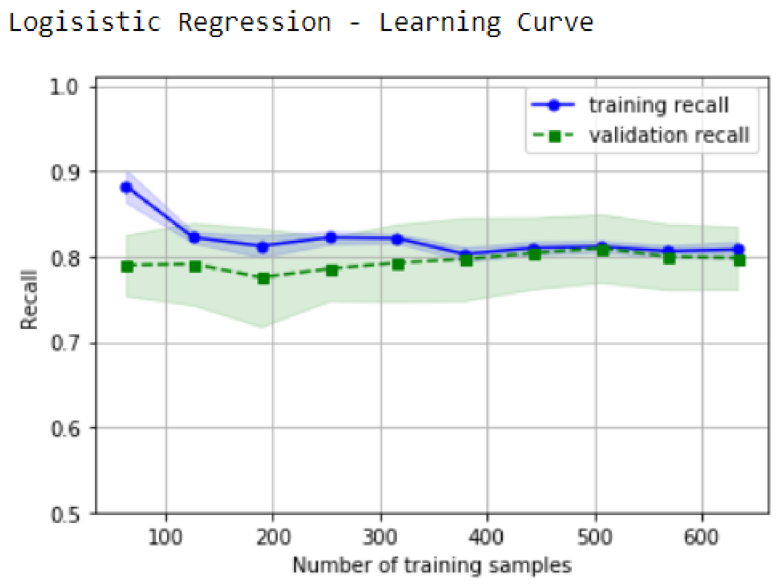
In the statistics table, count represents the sample size of each categorical variable. For those factors who's mean and, the standard deviation is high they influence the "Outcome" to affect most as compares to a variable whose gap is little.

The value of Standard deviation is 0.33 which is a low standard deviation indicates that most of the numbers are close to the mean 0.12.



```
[[128  26]
 [  5  18]]

               precision    recall  f1-score   support

   Outcome 0       0.96      0.83      0.89       154
   Outcome 1       0.41      0.78      0.54        23

    accuracy                           0.82       177
   macro avg       0.69      0.81      0.71       177
weighted avg       0.89      0.82      0.85       177


NestedCV Accuracy(weighted) :0.79 +/-0.06
NestedCV Precision(weighted) :0.88 +/-0.02
NestedCV Recall(weighted) :0.79 +/-0.06
```

- ➢ The figure reveals the confusion matrix of Logistic Regression in which precision determines how accurate our model is when it is closer to 100%.
- ➢ This model clarifies the division of the fluctuation between the quantities anticipated by the model and the value as opposed to the mean of the actual. This value is between 0 and 1.
- ➢ This model predicts a better negative variation in Utility because Recall for class 0 is 0.83 and class 0 has Recall 0.78, therefore it is easy to determine the positive and negative change in Utility.



The Learning Curve of Logistics Regression Model

The logistic regression learning curve compares the performance of a model on training and testing data over a varying number of training instances.

The gap between the training and validation recall is decreasing consistently which represents how well the model can generalize to new data.

We can determine from the graph that no issue of overfitting and underfitting which illustrates that good bias-variance trade-off that means model learning consistently and performance of model improves as the number of training points increases.

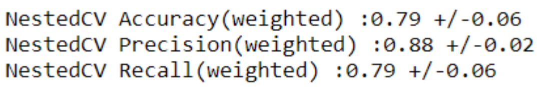
Here our y-axis is ‘Recall', not ‘SCORE', so the higher the Recall, the better the performance of the model.

Results

In the logistic regression, the value of accuracy, recall, f1-score is good that implies our algorithm was good to make a sensibly great prediction.

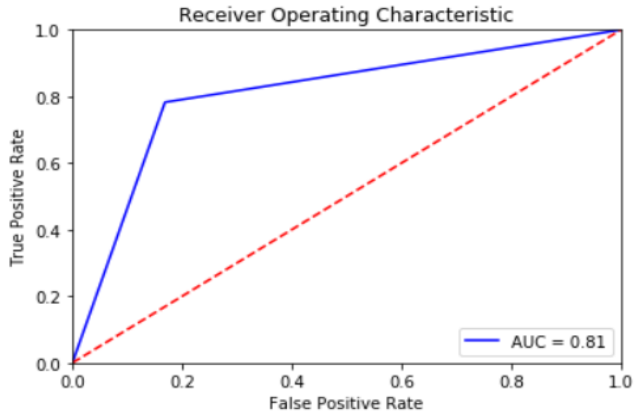
We should add some variable to create high accuracy in the model and should train the x and y better which builds good model.

Multiple data can also help to correlate between the data and variables, more inputs would help to create high level model.



```
NestedCV Accuracy(weighted) :0.79 +/-0.06
NestedCV Precision(weighted) :0.88 +/-0.02
NestedCV Recall(weighted) :0.79 +/-0.06
```

ROC Curve



NestedCV is used to train a model in which hyperparameters also need to be optimized. NestedCV estimates the generalization error of the underlying model and its hyperparameter search.

NestedCV for optimized logistic regression model is 0.79 +/- 0.06 which is often used to train a model in which hyperparameters also need to be optimized.

The recall of this model is 0.80 which demonstrates that the logistic regression model is good to predict positive and negative change in Utility.

Conclusion

SVM doesn't perform well, when we have large data set because the required training time is higher. 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 we need to replace NAs with mean to design a better predictive model and achieve high Recall. Our target audience will be Vice president of Starseed Medical Inc. 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.

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.

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.

References

Sam Plati_database overview- capstone (Data 2206)

Sam Plati_capstone)_reportexample (Data 2206