CSM 148 Project 3 Report

Haoyang Li 505522129

March 16, 2022

1 Background Information

The cannabis industry is a fast growing industry in the US, and California contributed almost one third of tax revenue from legal, adult-use cannabis sales in the US in 2021, and the total tax revenue from cannabis sales is over 3 billion US dollars.(3)

From 2018 to September 2021, the sales of cannabis in California increases averagely 50% per year, due to the COVID-19 pandemic in 2020 and 2021, the increase rate was not as high as 2018 and 2019, but the increased number was still high. (1)

The population of people consuming cannabis products is also high. According to the CDC statistics, 48.3 million Americans or 18% of total population used marijuana products at least once in 2019. (2)

Conducting further studies of the cannabis market can find the correlations between sales and consumers and discover consumers' behaviors to optimize the profit of cannabis companies.

2 Methodology

2.1 Part II

For part 2, I added a feature "current unit averages" that records the average sold unit for the last 4 months, and a feature "previous month units" that records the total units sold the month before.

The "current unit averages" feature takes the sum of total units of previous 4 months for the same brand then divide by 4, and the "previous month units" simply records the total units of previous month.

2.2 Part III

I dropped the columns "vs. Prior Period_x" and "vs. Prior Period_y" since they are temporary values from the last step. After iterating through the dataframe, I also found some brands with too low information that lacks too much data(more than 6 months), has no information about them(all nans) or have no products(ProdCount == 0).

I categorized the data set by the brands of the products.

I used median to impute all missing data because after inspecting the distribution of data, there are a number of outliers that are extremely high, and using mean of data would be inaccurate for further study.

I consider the way to use cannabis product important, so based on Category I of data set, I added 3 features "inhaleables", "ingestibles" and "topicals". I also counted the number of same brand and added a feature "ProdCount" to record this parameter.

2.3 Part VIII

I set max depth of my random forest to 8 and 3 estimators, and n_jobs to -1. The parameters of this grid search is not enough to make the result optimal, but would give a result with relatively high accuracy. If the amount of n_estimators parameter is too high, it would take forever for my computer to run.

2.4 Part IX

I used a random forest with max depth=10 and 10-fold cross validation as my best model.

3 Results

3.1 Result of Part 4

```
explained_variance:  0.5469
r2:  0.5465
MAE:  81654.9349
MSE:  19139470370.2771
RMSE:  138345.4747
```

Figure 1: Regression Result of Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Total Sales ($)   R-squared:                       0.573
Model:                            OLS   Adj. R-squared:                  0.554
Method:                 Least Squares   F-statistic:                     30.51
Date:                Wed, 16 Mar 2022   Prob (F-statistic):               0.00
Time:                        01:44:23   Log-Likelihood:            -2.7267e+05
No. Observations:               20642   AIC:                         5.471e+05
Df Residuals:                   19771   BIC:                         5.540e+05
Df Model:                         870
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          3.342e+04   3216.864     10.388      0.000    2.71e+04    3.97e+04
x2          3.963e+04   1271.662     31.166      0.000    3.71e+04    4.21e+04
x3         -8212.4137   1097.557     -7.482      0.000   -1.04e+04   -6061.109
x4          1.665e+04   1186.908     14.025      0.000    1.43e+04     1.9e+04
x5          2.608e+04   1443.543     18.067      0.000    2.33e+04    2.89e+04
x6         -9394.1532   3738.237     -2.513      0.012   -1.67e+04   -2066.895
x7           1.77e+04   2.88e+04      0.615      0.538   -3.87e+04    7.41e+04
x8          7.607e+04   3.02e+04      2.522      0.012    1.69e+04    1.35e+05
x9          1.194e+05   4.01e+04      2.979      0.003    4.08e+04    1.98e+05
x10          3.77e+05   2.65e+04     14.251      0.000    3.25e+05    4.29e+05
x11         2.907e+04    5.1e+04      0.570      0.568   -7.08e+04    1.29e+05
x12          5.91e+04    2.6e+04      2.276      0.023    8213.747     1.1e+05
x13         2.001e+05    2.5e+04      7.988      0.000    1.51e+05    2.49e+05
x14         2.396e+05   2.42e+04      9.891      0.000    1.92e+05    2.87e+05
x15         1.056e+05   3.27e+04      3.231      0.001    4.15e+04     1.7e+05
x16         1.272e+04    5.1e+04      0.250      0.803   -8.71e+04    1.13e+05
x17        -5745.5995   2.23e+04     -0.258      0.797   -4.95e+04     3.8e+04
x18         5.563e+04   3.48e+04      1.598      0.110   -1.26e+04    1.24e+05
x19         5.589e+05   2.65e+04     21.128      0.000    5.07e+05    6.11e+05
x20         9.238e+04   3.72e+04      2.481      0.013    1.94e+04    1.65e+05
x21         2.576e+04   4.07e+04      0.634      0.526   -5.39e+04    1.05e+05
x22         4.361e+05   2.26e+04     19.256      0.000    3.92e+05     4.8e+05
x23         3.279e+04   4.26e+04      0.769      0.442   -5.08e+04    1.16e+05
x24         1.099e+05    2.6e+04      4.235      0.000    5.91e+04    1.61e+05
x25         9.249e+04   4.74e+04      1.950      0.051    -460.744    1.85e+05
```

Figure 2: Statistics of Linear Regression

x846	9.272e+04	3.09e+04	2.997	0.003	3.21e+04	1.53e+05
x847	2.311e+04	4.77e+04	0.485	0.628	-7.03e+04	1.17e+05
x848	4.73e+04	4.77e+04	0.992	0.321	-4.61e+04	1.41e+05
x849	4.885e+04	2.88e+04	1.699	0.089	-7512.931	1.05e+05
x850	8.625e+04	2.8e+04	3.078	0.002	3.13e+04	1.41e+05
x851	9573.8845	4.77e+04	0.201	0.841	-8.39e+04	1.03e+05
x852	2.431e+05	2.28e+04	10.662	0.000	1.98e+05	2.88e+05
x853	2.837e+05	3.02e+04	9.396	0.000	2.25e+05	3.43e+05
x854	5.936e+04	4.49e+04	1.321	0.187	-2.87e+04	1.47e+05
x855	1.984e+04	3.74e+04	0.530	0.596	-5.35e+04	9.31e+04
x856	2.515e+05	2.22e+04	11.341	0.000	2.08e+05	2.95e+05
x857	3.196e+04	2.81e+04	1.137	0.256	-2.32e+04	8.71e+04
x858	5.537e+04	4.05e+04	1.366	0.172	-2.41e+04	1.35e+05
x859	9.981e+04	2.22e+04	4.502	0.000	5.64e+04	1.43e+05
x860	1.703e+05	2.28e+04	7.460	0.000	1.26e+05	2.15e+05
x861	2.223e+04	4.07e+04	0.547	0.584	-5.74e+04	1.02e+05
x862	1.977e+04	3.48e+04	0.568	0.570	-4.85e+04	8.8e+04
x863	1.512e+05	3.88e+04	3.895	0.000	7.51e+04	2.27e+05
x864	4.704e+05	2.58e+04	18.257	0.000	4.2e+05	5.21e+05
x865	4.715e+04	2.42e+04	1.946	0.052	-348.356	9.46e+04
x866	8.995e+04	2.54e+04	3.537	0.000	4.01e+04	1.4e+05
x867	2.484e+04	3.74e+04	0.664	0.507	-4.85e+04	9.81e+04
x868	2.114e+04	4.26e+04	0.496	0.620	-6.24e+04	1.05e+05
x869	4.284e+05	2.21e+04	19.393	0.000	3.85e+05	4.72e+05
x870	2.797e+04	4.26e+04	0.656	0.512	-5.56e+04	1.12e+05
x871	6.702e+04	2.22e+04	3.021	0.003	2.35e+04	1.11e+05
x872	1.009e+05	2.22e+04	4.556	0.000	5.75e+04	1.44e+05
x873	1.654e+04	5.1e+04	0.325	0.745	-8.33e+04	1.16e+05
x874	2.49e+04	2.6e+04	0.959	0.337	-2.6e+04	7.58e+04
x875	1.005e+05	5.07e+04	1.983	0.047	1148.434	2e+05

Omnibus:	5418.250	Durbin-Watson:	0.876
Prob(Omnibus):	0.000	Jarque-Bera (JB):	33483.666
Skew:	1.117	Prob(JB):	0.00
Kurtosis:	8.826	Cond. No.	8.54e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 6.44e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Figure 3: Statistics of Linear Regression(continued)

Although the R^2 value is above 0, many coefficients have extremely high or low values, the extremely high values are caused by the very high values of total sales, and the very low values means that the variables are insignificant; also, the mae, mse and rmse are all very high.

3.2 PCA

```
explained_variance:  0.1579
r2:  0.1575
MAE:  133255.0419
MSE:  35558169862.5784
RMSE:  188568.7404
                            OLS Regression Results
=======================================================================================
Dep. Variable:        Total Sales ($)   R-squared (uncentered):                   0.097
Model:                            OLS   Adj. R-squared (uncentered):              0.096
Method:                 Least Squares   F-statistic:                              165.5
Date:                Wed, 16 Mar 2022   Prob (F-statistic):                   7.78e-135
Time:                        02:49:50   Log-Likelihood:                         -85795.
No. Observations:                6193   AIC:                                  1.716e+05
Df Residuals:                    6189   BIC:                                  1.716e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          5.308e+04   2091.598     25.377      0.000     4.9e+04    5.72e+04
x2         -2120.0062   2505.878     -0.846      0.398   -7032.398    2792.385
x3          6355.8391   3367.168      1.888      0.059    -244.980     1.3e+04
x4          2.046e+04   3793.097      5.394      0.000     1.3e+04    2.79e+04
==============================================================================
Omnibus:                     2057.472   Durbin-Watson:                   1.146
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7007.834
Skew:                           1.676   Prob(JB):                         0.00
Kurtosis:                       6.990   Cond. No.                         1.82
==============================================================================
```

Figure 4: Statistics of PCA

The R^2 of PCA optimized model using logistic regression performs slightly better than an average predictor. The p-value of x_2 is high, which means it is less likely to be a highly influential variable.

3.3 Ensemble Methods

```
explained_variance:  0.6091
r2:  0.6083
MAE:  75435.1771
MSE:  16532587135.636
RMSE:  128579.1085
```

Figure 5: Statistics of Random Forest

With an R^2 value of 0.618, thus it performs pretty well. Although the MAE, MSE and RMSE are still high, it performs slightly better than the linear regression model in part IV.

3.4 Cross Validation

```
building models...
generating results...
done!
Linear Regression Accuracy: 52.20%
Random Forest Accuracy: 62.65%
```

Figure 6: Result of 2 Cross Validations

```
explained_variance:  0.6481
r2:  0.6472
MAE:  69372.605
MSE:  14891771383.4618
RMSE:  122031.8458
```

Figure 7: Result of Random Forest Cross Validations

```
explained_variance:  0.5469
r2:  0.5465
MAE:  81654.9349
MSE:  19139470370.2771
RMSE:  138345.4747
```

Figure 8: Result of Linear Regression Cross Validations

3.5 Grid Search

After applying grid search to get a set of optimized parameter for cross validation for linear regression, the results are as below:

```
print('Negative mae:', gridResult.best_score_)
print('Hyperparameters:', gridResult.best_params_)

Negative mae: -83932.5151672646
Hyperparameters: {'max_depth': 8, 'n_estimators': 100, 'n_jobs': -1}
```

Figure 9: Result of Linear Regression Cross Validations

After applying 10-fold cross validation, the performance of random forest increased a lot, and the performance of linear regression stayed the same. Thus in this case, applying cross-validation has more effect on random forest than on linear regression.

3.6 Best Model Performance

This is the performance of my random forest model with $max\_depth = 10$ and a 10-fold cross validation.

```
explained_variance:  0.6481
r2:  0.6472
MAE:  69372.605
MSE:  14891771383.4618
RMSE:  122031.8458
```

Figure 10: Result of My Random Forest Model

4 Discussion

Using PCA to optimize the performance is problematic for this problem. It's true that there are a lot of variables that determines the sales of cannabis products, but they have similarly low influence and it's hard to say if there are some of them weigh significantly higher than others.

According to the result of my approaches, a random forest model best simulates and predicts the total sales of a specific brand of cannabis. With a better computer, analyzers should find it easy to fit a better set of parameters for cross validation and grid search.

However, to optimize the predictor, more variables should be considered, including weather that influences the harvest of weed which would influence the cost of producing correlated products, general price level of supermarket goods that decides if there are enough money for people to spend on cannabis products, and even important news that influence people's everyday's life and would influence their decision of whether purchasing some cannabis products.

References

[1] Project, M. P. (n.d.). *Cannabis tax revenue in states that regulate cannabis for adult use.* MPP. Retrieved March 15, 2022, from https://www.mpp.org/issues/legalization/cannabis-tax-revenue-states-regulate-cannabis-adult-use/

[2] Centers for Disease Control and Prevention. (2021, June 8). *Data and statistics* Centers for Disease Control and Prevention. Retrieved March 15, 2022, from https://www.cdc.gov/marijuana/data-statistics.htm

[3] Ramos, E. (2022, January 21). *Recreational marijuana sales showered states with cash in 2021* NBC News. https://www.nbcnews.com/news/us-news/recreational-marijuana-sales-showered-states-cash-2021-n1287861