Cannabis Brand Forecasting Report

Andrew Bennecke and Daniel Frees
CS 148

Table of Contents

1. Executive Summary

The cannabis market has seen record growth over the past few years and is projected to continue this growth throughout the decade. One key way that other industries have optimized their growth over the past several years has been through predictive and analytical machine learning models. The cannabis market is no exception to this practice and stands to accelerate its already rapid growth by exploring the vast quantities of pre-existing data relating to their industry. In this report, we explore a subset of this vast data and extract actionable insights.

First, several time-series and brand-level features were augmented in order to enhance the scope of the model. Several of the time series features include past sales data, rolling averages of units moved, brand market share, and brand length in market. Several of the brand-level features include product categorizations such as whether brands sell THC-containing products, and whether brands sell ingestible cannabis products. We then dropped all features with a high collinearity to the target feature *Total Sales ($)* because one of our primary goals was to identify uncommon features which may be strong predictors of *Total Sales ($)*.

Once the data was cleaned, augmented, and imputed, 3 different models were explored. The first was a basic Linear Regression Model. The latter 2 methods employed an ensemble approach. The first of these was eXtreme Gradient Boosting Regression (XGB) and the second of these was Random Forest Regression (RFR). The hyperparameters of the XGB model was optimized using GridSearchCV, and the RFR model was selected and tuned via manual exploration. After identifying optimal sets of hyperparameters, each model was trained, tested, and K-fold cross-validated. To extract insights, confidence intervals were generated for linear regression coefficients, and cross-validated feature importances were generated for the ensemble models.

Cross-Validated Results: Model performance was extremely poor for the linear regression model (R^2=0.2639, MAE/μ(sales)=0.9996). The XGB model was great at explaining variance in *Total Sales ($),* but still not particularly precise (R^2=0.8897, MAE/μ(sales)=0.4437). The RFR model was the best at explaining variance in *Total Sales ($),* and was the most precise (R^2=0.8961, MAE/μ(sales)=0.3695).

Since model performance was so low for linear regression, inferences were made based on the XGB and RFR models, combined with Pearson's correlation coefficients between features and *Total Sales ($)*. The most significant positive predictor was *NumProducts*, indicating that some combination of product variety and brand size is a powerful indicator of *Total Sales ($).* Average retail price and recent market share growth were also consistently important predictors, suggesting that retail pricing and brand momentum are important factors driving monthly sales.

In conclusion, *Total Sales ($)* for future months of cannabis sales can be predicted with moderate accuracy even after removing previous month sales and other highly correlated features. Upon generating models without these features, numerous avenues for further brand research become obvious.

2. Background / Introduction

As of October 2021, 19 states have fully legalized marijuana use, 26 states have to some extent legalized medical marijuana use, and 5 states have legalized the use of cannabidiol (CBD)[1]. On a consumer level, 73% of adults in the U.S. are open to cannabis consumption[1]. With this partial or total legalization of marijuana across all states in the U.S, the cannabis industry has seen incredible growth over the past several years. The majority of this growth over the course of 2020 has been due to a very large increase in the sale of edibles. The global edible market (mostly dominated by the U.S) in particular saw a growth of 24% from 2019 to 2020 while the global cannabis market (mostly dominated by the U.S.) saw a growth of ~28%[2]. Over the next 5 years, the U.S. Legal Cannabis Market is expected to grow from 25 billion $USD to 47 billion $USD[1].

Within the revenue-dominating U.S. market, California is one its largest contributors. Within this state, the top sales categories in descending order are flower, concentrates, edibles, sublinguals, topicals, and pre-rolled[3]. These 6 categories constitute 1.9 billion $USD of the ~3.5 billion $USD cannabis market size of California. These large contributors drive some of the later decisions we make during model development and analysis.

Aside from the growth of these existing cannabis markets, their novelty allows for rapid innovations in new methods of consumption. Originally, the primary products were flower-based or edible-based. However, there have been many recent developments in cannabis-based beverages, vapors, waxes, and consumables over the past two years. In order to develop and sell all of these products, every brand within the industry must fulfill one or more aspects of the licensing pipeline. This process consists of 4 steps: (1) Cultivation, (2) Manufacturing, (3) Distribution, and (4) Retail[2]. At every step of this process, a plethora of computational methods could be employed to optimize results. For example, Machine Learning (ML) could be used to optimize growth conditions for the cultivation of the marijuana plant. Artificial Intelligence (AI) could be used for Supply Chain optimization at the manufacturer and distributor steps. Finally, Data Science methods can be used to identify which aspects of a product or vendor correlate with higher sales.

In the following report, we extract time series cannabis sales data and brand-level product data from pre-existing datasets. Using this data, we explore several machine learning models for extracting actionable insight, which may be used to improve overall revenues for new or existing companies within the cannabis retail market. Furthermore, we identify important predictors whose precise effect remains unclear, and suggest future directions to elucidate the exact sales impact of these predictors.

---

[1] BDSA's Fall 2021 Cannabis Market Forecast Update
[2] BDSA 2021 State of Cannabis (by Jessica Lukas)
[3] BDSA Data Deeper: California

3. Methodology

Time Series Feature Extraction

In order to extract features using the time series aspects of the data, we began extracting different features from a single brand. We then expanded to all brands contained within the dataset.

The first feature we extracted was *Year*, which describes the year of the corresponding *Total Sales ($)*. The second feature we extracted was *Quarter*, which describes the fiscal quarter of the corresponding *Total Sales ($)*. After performing some industry research, we determined that the Cannabis Industry splits its fiscal quarters as follows: (Q1) Jan-Mar, (Q2) Apr-Jun, (Q3) Jul-Sep, and (Q4) Oct-Dec. The third feature we extracted was *length_in_market*, which describes the total length of time (in days) that a brand has been present since its first instance in the dataset.

Next, we utilized adjacent months of sales data to extract time series features. Via this methodology, the first feature we extracted was *Previous Month Sales*, which corresponds to the individual brand's *Total Sales ($)* from the previous month. If this was the brand's first month in the dataset, this value was later imputed as 0. The second feature we extracted was *Rolling Average Sales*, which describes the average *Total Sales ($)* each month from the past 3 months. The third feature we extracted was *Market Share (%)*, which describes the brands total percentage of the market within the dataset. This value was calculated by using the current *Total Sales ($)* data of a date and dividing it by the *Total Sales ($)* across all brands on that date. Since *Market Share (%)* uses current sales data, we utilized this feature to extract 3 other features corresponding to past *Market Share (%)* data. These 3 features were *Previous Month Market Share*, *Rolling Average Market Share*, and *Market Share Growth of Prev. 2 Months*.

Following this, we utilized adjacent unit movement data (indicating units sold in a given month) to extract time series features corresponding to units moved from past months. We extracted *Previous Month Units Moved* and *Rolling Average Units Moved* in the similar way that we extracted *Previous Month Sales* and *Rolling Average Sales*.

Brand Level Feature Extraction

In order to extract detailed features about each brand's product line we iteratively developed feature extraction code on a single brand, then expanded this code to extract similar information from all brands in the dataset.

In order to determine which features to extract, we first considered all of the features within *BrandDetails.csv* and bucketed them accordingly. The following features were ignored due to:

1) Uniformity across all instances of the dataset: *State* and *Channel*

2) Large majority of null values: *Category L5*, *Flavor*, *Item Weight*, *Items Per Pack*, and *Pax Filter*.
3) Too many unique non-numeric values: *Category L3*, *Product Description*, *Strain*
4) Collinearities with *Total Sales ($)*: *Total Units*
5) Redundancy due to added features: *Contains CBD*, *$5 Price Increment*
6) Readily apparent lack of descriptive value: *Mood Effect*, *Generic Vendor*, *Generic Items*

Based on Category L1, we determined which brands sold products classified as 'inhalables' and which brands sold products classified as 'ingestibles'. We then created boolean features accordingly. We also determined whether each of these brands sold a majority of Inhalables or Ingestibles (*Product Majority*).

Based on Category L2, we determined which brands sold products classified as 'Concentrates' and which brands sold products classified as 'Pre-Rolled' or 'Flower' accordingly. We then created boolean features accordingly.

Based on Category L4, we determined which brands sold products classified as 'Vape Cartridge' and created a boolean feature accordingly.

Based on 'Total THC' and 'Total CBD', we determined which brands sold products containing THC and which brands sold products containing CBD. We then created boolean features accordingly.

Lastly, we counted the number of products sold by each individual brand, by counting the length of each brand's product list. This product count was then linked to each brand to generate a new feature, *NumProducts.*

Drops and Imputation

Following feature extraction and data visualization, we elected to remove the following features due to:

1) Collinearity with *Total Sales ($)*: *3 Prev. Month Units Moved Avg., Total Units, Units Moved vs. Prior Period, Units Moved Previous Month, Market Share (%), Previous Month (Market Share), Rolling Average (Market Share)*
2) Redundancy/Multicollinearity: *Sells CBD, Product Majority*
3) Lack of usability aside from the extraction process itself: *Brands, Months*

Note that time series information from previous time points (such as *Units Moved Previous Month)* would be valid inclusions in a sales forecaster if our goal was solely to predict Total Sales with the most precision possible. However, our goal was to draw meaningful insights from the data, and therefore we avoided retaining features which might lead the model to simply predict a Brand's monthly sales to be nearly the same as their previous month's sales.

Due to the lack of ordinality and the small quantity of unique values, we one-hot encoded *quarter* into 4 dummy variables corresponding to each fiscal quarter.

The following features were imputed by filling *NaN* values as one can reasonably assume that a 'non-effect' value for these features should be 0 (they correspond to changes/growth): *Average 3 Month Units Growth*, *ARP vs Prior Period*, *ARP Increase Prev 2 Months ($)*, and *Market Share Growth (% of Market) over Previous 2 Months*.

Since *ARP* was not an augmented feature and it had a right-skewed distribution, we imputed using the median of all ARP values across all brands and dates.

All numerical training data was scaled using *StandardScaler()* to prevent the overrepresentation of specific features with noticeably larger magnitudes (in case kNN or another model affected by variance were to prove useful through experimentation).

Ensemble Methods

XGBoost

We decided to use the eXtreme Gradient Boosting Regressor (*xgboost.XGBRegressor()*) for our first ensemble method of *Total Sales ($)* prediction. In order to determine which parameters would result in the highest model performance, we started by performing some manual hyperparameter tuning. Through this manual approach, we determined that the following non-default parameters resulted in higher performance than the basic approach:

- Objective Function: Squared Error
- Booster: gbtree
- Max_depth: 4

We also considered adding the *Gamma* hyperparameter if overfitting appeared to be a problem through later analysis. Further tuning was performed using GridSearchCV.

RF Regression

For our experimental approach, we elected to use the ensemble method Random Forest Regressor (*sklearn.ensmble.RandomForestRegressor()*). In order to determine which parameters would result in the highest model performance, we again performed some manual hyperparameter tuning. Through this manual approach, we determined that the following non-default parameters resulted in higher performance than the basic approach:

- n_estimators: 180
- max_features: 16
- oob_score: True

Cross-Validation Approaches

Performance metrics and Feature Importance

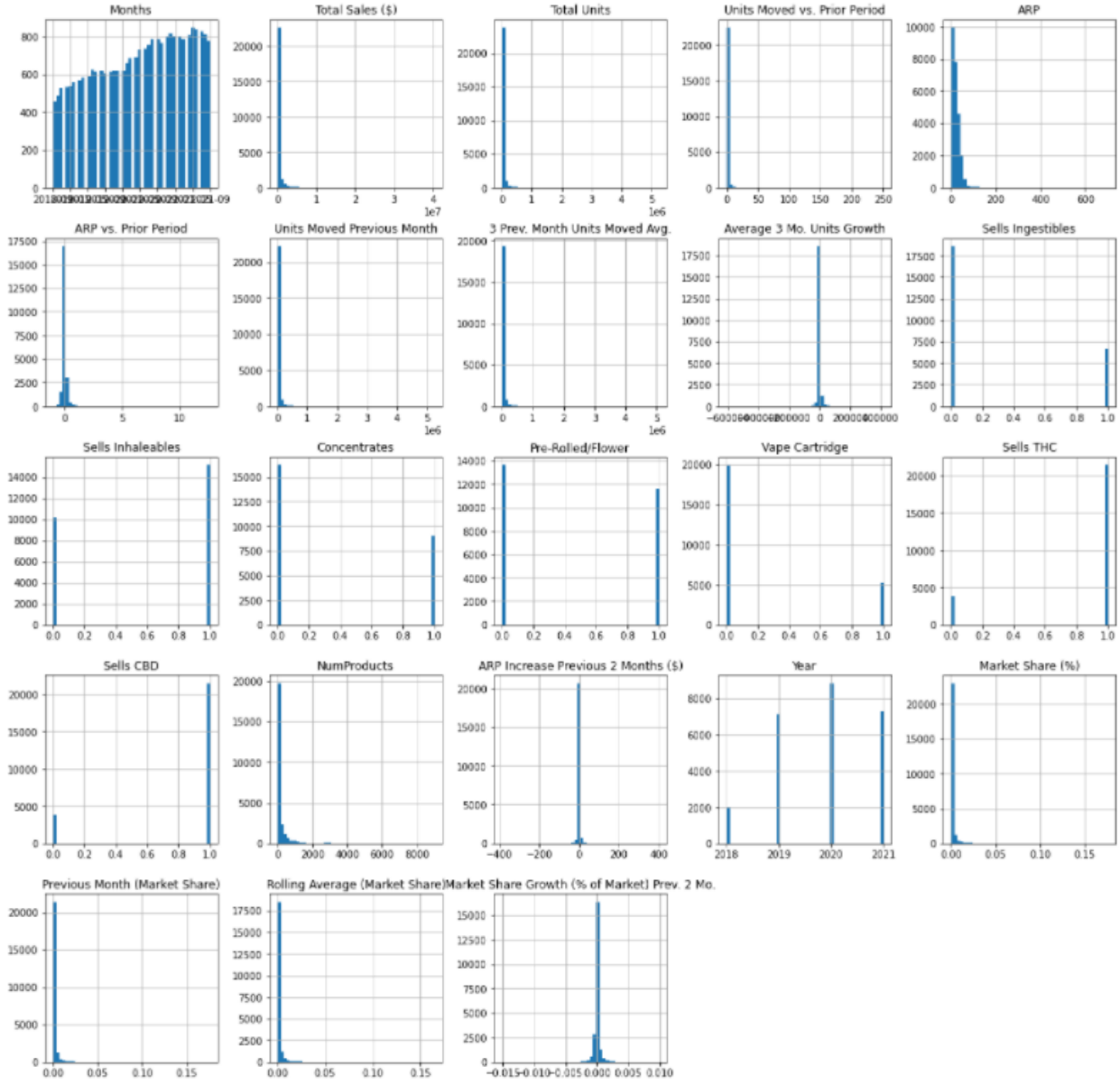
In order to perform a Stratified K-Fold Split, we decided to augment a new feature: *Total Sales Quartile*. Using this feature, we binned each instance of the dataset into their quartile based on the distribution of *Total Sales ($).* After updating the data with this new feature, we performed a Stratified 10-fold split on the dataset, ensuring an even distribution of *Total Sales Quartile*. Along with MAE and R^2, we also performed cross-validation across the feature importances of each feature from the ensemble methods, and generated standard deviations for these feature importances.

Hyperparameter Optimization (GridSearchCV)

In order to determine the optimal combination of hyperparameters to use for the XGB model, we employed the Grid Search Cross Validation (*sklearn.model_selection.GridSearchCV()*) approach.

4. Results

Basic Statistics and Data Exploration



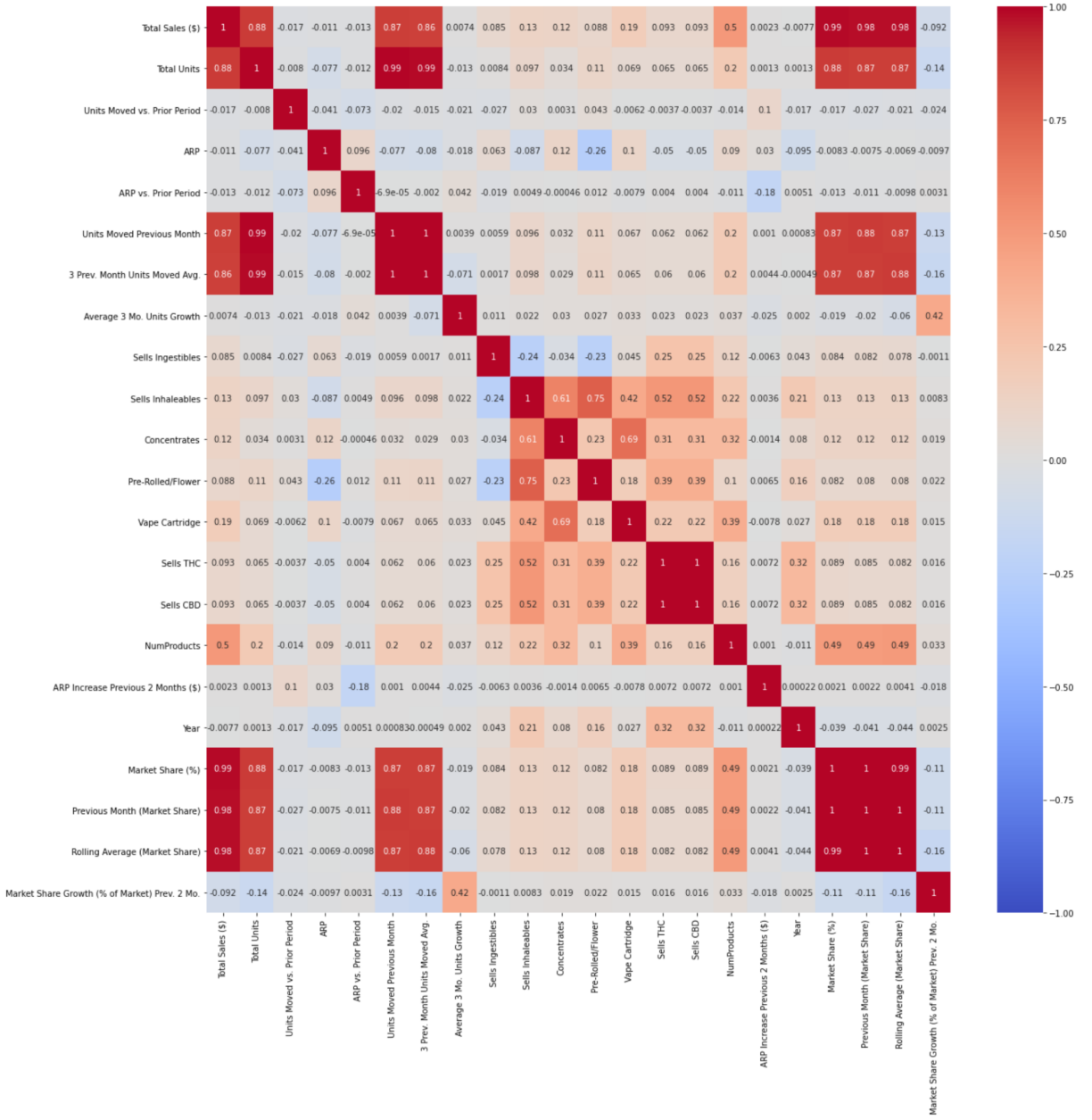
Histograms of all original and augmented features

From the above figure, we can make the following observations for each feature:

1) ‘Months’, which was actually a feature filled with datetimes, is distributed unevenly with a bias towards more recent dates.
2) Extremely right-skewed distribution: *Total Sales ($), Total Units, Units Moved vs. Prior Period, Units Moved Previous Month, 3 Prev. Month Units Moved Avg., NumProducts,*

Market Share (%), Previous Month (Market Share), Rolling Average (Market Share), Market Share Growth (%) Prev. 2 Mo.

3) Right-skewed distribution: *ARP*
4) Normal distribution with extremely low variance: *ARP vs. Prior Period, Average 3 Mo. Units Growth, ARP Increase Previous 2 Months ($)*
5) Bernoulli distribution (Majority 0): *Sells Ingestibles, Concentrates, Pre-Rolled/Flower, Vape Cartridge*
6) Bernoulli distribution (Majority 1): *Sells Inhalables, Sells THC, Sells CBD*
7) Discrete normal distribution: *Year*



Heatmap of correlation matrix across all features (correlation values correspond to Pearson's correlation coefficients)

From the above figure, we can make the following observations for all features with respect to the target feature, *Total Sales ($)*:

1) Highly positive correlation: *Total Units, Units Moved Previous Month, 3 Prev. Months Units Moved Avg, Market Share (%), Previous Month (Market Share), Rolling Average (Market Share)*
2) Positive Correlation: *NumProducts*
3) Slightly positive correlation: *Sells Ingestibles, Sells Inhalables, Concentrates, Pre-Rolled/Flower, Vape Cartridge, Sells THC, Sells CBD*
4) Virtually no correlation: *Units Moves vs. Prior Period, ARP, ARP vs. Prior Period, Average 3 Mo. Units Growth, ARP Increase Previous 2 Months (%), Year*
5) Slightly negative correlation: *Market Share Growth (% of Market) Prev. 2 Mo*

Principal Component Analysis (PCA)

In order to simplify the complexity of the data, we performed a dimensionality reduction that would retain 95% of the variance of the data. This resulted in the loss of 4 features from the pre-reduced data. We trained and tested the LinReg and XGB models using both the pre-reduced and reduced- datasets. The results are discussed in their respective sections.

Basic Linear Regression Model

The first model we considered was a linear regression model (*sklearn.linear_model.LinearRegression*) with no hyperparameter tuning. From this model, we obtained the following results from the training data:

1) Mean Absolute Error / Average Total Sales: 1.011
2) R^2: 0.2818

We obtained the following results from the testing data:

1) Mean Absolute Error / Average Total Sales: 0.9778
2) R^2: 0.2261

The features with the 3 highest positive coefficients are:

1) *NumProducts*: 813,077
2) *Sells Inhalables*: 185,307
3) *Length in Market (Days)*: 143,002

Lastly, the features with the 3 lowest negative coefficients are:

1) *Market Share Growth (% of Market) Prev. 2 Mo.*: -210,525
2) *Concentrates*: -165,699
3) *Year*: -129,595

We also ran a linear regression model from the statsmodels API (*statsmodels.api.OLS)* to generate confidence intervals for all linear regression parameters.

All of the following features had coefficients with p-values < 0.002 and effect directionality confirmed by 95% confidence intervals:

NumProducts had the largest positive effect on total sales. *ARP* reliably expressed a negative coefficient, indicating that *ARP* is negatively correlated with *Total Sales ($). Sells Ingestibles* and *Sells Inhaleables* were both positively correlated with sales. On the other hand, the sale of *Concentrates* had a negative effect on sales. *Length in Market* had a reliably large effect on sales. *Year* was negatively correlated with *Total Sales ($),* as was *quarter_4* (The fourth fiscal quarter of the year).

PCA

Training and testing the Basic Linear Regression model on the dimensionality reduced data resulted in the following metrics for training data:

1) Mean Absolute Error / Average Total Sales: 0.9907
2) R^2: 0.2791

We obtained the following results from the testing data:

1) Mean Absolute Error / Average Total Sales: 0.9535
2) R^2: 0.2251

10-Fold Cross-Validation

Utilizing the Stratified 10-Fold splitting of the data along the *Total Sales Quartile* feature, we obtained the following results on the test data:

1) Average Mean Absolute Error / Average Total Sales: 0.9996
2) Average R^2: 0.2639

Ensemble Method I: XGBoost

The second model we considered was the *XGBRegressor()*. The manual hyperparameter tuning resulted in the aforementioned specifications (see Methods). From this model, we obtained the following results for the training data:

1) Mean Absolute Error / Average Total Sales: 0.3295
2) R^2: 0.9717

We obtained the following results for the testing data:

1) Mean Absolute Error / Average Total Sales: 0.4260
2) R^2: 0.9119

PCA

Training and testing the modified XGB Regressor model on the dimensionality reduced data resulted in the following metrics for training data:

1) Mean Absolute Error / Average Total Sales: 0.4235
2) R^2: 0.9539

We obtained the following results for the testing data:

1) Mean Absolute Error / Average Total Sales: 0.6140
2) R^2: 0.6653

10-Fold Cross Validation

Utilizing the Stratified 10-Fold splitting of the data along the *Total Sales Quartile* feature, we obtained the following results on the test data:

1) Average Mean Absolute Error / Average Total Sales: 0.4437
2) Average R^2: 0.8897

Through the cross-validation, we also considered the importance of each feature, based on the ‘gain’ metric. See the following table.

	Feature	Importance (Based on Gain)	StDev Relative Importance:
0	Concentrates	6.537618	8.698764
1	NumProducts	3.915346	0.466327
2	ARP	2.031717	0.192815
3	Market Share Growth (% of Market) Prev. 2 Mo.	0.950243	0.143151
4	Sells Ingestibles	0.860990	0.209412
5	Average 3 Mo. Units Growth	0.622520	0.062877
6	Pre-Rolled/Flower	0.611277	0.188295
7	Sells Inhaleables	0.593081	0.249391
8	Length In Market (Days)	0.592251	0.103603
9	Vape Cartridge	0.420146	0.211440
10	ARP vs. Prior Period	0.223261	0.067125
11	Year	0.191790	0.097291
12	ARP Increase Previous 2 Months ($)	0.171984	0.030177
13	quarter_1	0.097688	0.081879
14	quarter_3	0.071835	0.040681
15	quarter_2	0.056747	0.020981
16	quarter_4	0.051064	0.027854
17	Sells THC	0.000442	0.000000

Cross-Validated Feature Importance Based on Gain of XGBoost

Based on the XGB model, we note that the following 3 features have the highest importance:

1) Concentrates: 6.53x > average feature importance, 8.69x SD
2) NumProducts: 3.92x > average feature importance, 0.47x SD
3) ARP: 2.03x > average feature importance, 0.19x SD

GridSearchCV

We first considered the following sets of hyperparameter combinations:

1) Objective Function: SquaredError
2) Lambda: [0, 0.01, 0.1]
3) Alpha: [0, 0.01, 0.1]
4) Max_depth: [3, 4, 5]
5) Booster:

a) gbtree: sampling_method = [uniform, gradient_based]
b) gblinear: updater = [shotgun, coord_descent]

From these hyperparameter combinations, we determined that the optimal solution for both the explained_variance and R^2 metrics, with a score of 0.897, was:

{ ‘alpha’: 0, ‘booster’: gbtree, ‘lambda’: 0.1, ‘max_depth’: 5, ‘objective’: ‘reg:squarederror’, ‘sampling_method’: uniform, ‘random_state’: 42 }

Ensemble Method II: Random Forest Regressor

The third and final model we considered was the *RandomForestRegressor()*. The manual hyperparameter tuning resulted in the aforementioned specifications (see Methods). From this model, we obtained the following results for the training data:

1) Mean Absolute Error: $42,187.74
2) R^2: 0.988

From this model, we obtained the following results for the testing data:

1) Mean Absolute Error: $108,391.89
2) R^2: 0.9376

10-Fold Cross-Validation

After performing a Stratified Split along the *Total Sales Quartile* feature and a 10-Fold Cross Validation, we obtained the following metric results:

1) Average Mean Absolute Error / Average Total Sales: 0.3695
2) Average R^2: 0.8961

Feature Importance Cross Validation

	Feature	Relative Importance	StDev Relative Importance:
0	NumProducts	7.859695	0.422763
1	ARP	2.590555	0.292484
2	Market Share Growth (% of Market) Prev. 2 Mo.	2.257694	0.393573
3	Average 3 Mo. Units Growth	2.224174	0.461592
4	Concentrates	0.889798	0.128509
5	Vape Cartridge	0.614890	0.094516
6	Length In Market (Days)	0.602146	0.047006
7	Sells Ingestibles	0.246258	0.022626
8	ARP Increase Previous 2 Months ($)	0.155040	0.010432
9	ARP vs. Prior Period	0.151704	0.009787
10	Sells Inhaleables	0.139838	0.017669
11	Year	0.120879	0.010904
12	Pre-Rolled/Flower	0.071380	0.006410
13	quarter_3	0.023955	0.002823
14	quarter_4	0.018284	0.003615
15	quarter_1	0.017508	0.002068
16	quarter_2	0.016000	0.001635
17	Sells THC	0.000203	0.000038

Cross-Validated Feature Importance Based on Gain of RF Regression

Based on the RF Regressor model, we see that the following 3 features have the highest importance:

1) NumProducts: 7.85x > average feature importance, 0.42x SD
2) ARP: 2.59x > average feature importance, 0.29x SD
3) Market Share Growth (%) Prev. 2 Mo.: 2.25x > average feature importance, 0.39x SD

5. Discussion

PCA

95% variance PCA decreased model performance for both extreme gradient boosting and linear regression, across both MAE and R^2 metrics. This indicates that the rigorous multicollinearity removal and feature selection process was successful at removing redundant information, and dimensionality reduction is unnecessary for the pipelined data produced for this analysis.

GridSearchCV

The agreement of the explained_variance (biased variance) and R^2 metrics in determining ideal XGB model hyperparameters is nontrivial. This suggests that residuals for the XGB model are unbiased, centering around 0.

Basic Linear Regression Model

Although some interesting correlations for feature importance can be drawn from this model, the extremely low R^2 values for both training and testing data indicate that this model does a poor job of explaining the variance in *Total Sales ($)*.

After performing the 10-Fold cross validation, we only see a slight improvement of the R^2 value for the testing data, but this improvement does not change the status of Linear Regression as a poor predictive model for *Total Sales ($)*.

Linear regression enables very simple parameter interpretation, both in terms of the magnitude and directionality of each feature's effect on total sales. However, because linear regression was so poor at explaining the variance in *Total Sales ($)*, it would be irresponsible to make conclusions based on the parameter importance results. It is better to consider Pearson's correlation values in tandem with the cross-validated parameter importance results calculated for the ensemble models. See below.

Ensemble Method I: XGBoost

XGBoost was an excellent model for explaining *Total Sales ($)* variance with the given features. Therefore, it is reasonable to dive deeper into the feature importance results for actionable business takeaways. The following discussion is based on cross-validated performance and feature importance for the XGBoost model with parameters: *objective="reg:squarederror", booster = 'gbtree', max_depth = 4*. The cross-validated R^2 value for this model was 0.8897, indicating that the model was able to explain approximately 89% of the variance in *Total Sales ($).*

Considering both the cross-validated relative importance of features, as well as the standard deviation of these relative importances, the most consistently critical features for predicting *Total*

Sales ($) appear to be *NumProducts* and *ARP*. *Concentrates* has the highest relative importance value, but is inconsistently useful across different cross-validation splits, suggesting that it is relevant to only a subset of the brands and timepoints available in the data.

Based on the correlation analysis run before modeling, *NumProducts* is highly positively correlated (r=0.5) with *Total Sales ($)*. This suggests that a wider range of products is an important positive predictor of brand sales success. While this may simply result from the fact that larger cannabis companies are likely to sell more products, it may also indicate that customers appreciate a greater variety of products to choose from. As such, further research should be conducted with focus groups to determine whether Cookies' customers are interested in seeing a wider variety of products in stores.

ARP is not highly correlated (r=-0.011) with *Total Sales ($)*, yet it is the third most important feature based on gain. This suggests that ARP must be considered alongside other brand details and market information to determine the directionality of its effect. The current model is not sufficient to draw conclusions about ideal retail pricing, but it does indicate that intelligent pricing is extraordinarily important to ensure overall brand sales success.

It is reasonable to separately judge the importance of binary features, as 'gain'-based importance (the importance metric used for this analysis) is biased to inflate the importance of continuous and high-cardinality features, leaving binary features with lower importance values. Among the binary features, the most important was whether a brand *Sells Ingestibles.* This was moderately important, with an average relative importance of 0.861 and a standard deviation of 0.21. Second most important among the binary features was whether a brand sells pre-rolled joints or flower products (*Pre-Rolled/Flower)*.

Both of these binary features are slightly positively correlated with Total Sales ($), with Pearson's coefficients of 0.085 and 0.088 respectively. Unfortunately, the correlation values are too close to 0 to draw concrete conclusions about the directionality of effect when incorporating edibles and flower products into a product line.

Ensemble Method II: Random Forest Regressor

Among all ML models evaluated throughout this project, the Random Forest Regressor was best able to explain variance in *Total Sales ($),* with a cross-validated R^2 value of 0.8961. Thus, this model's feature importances can reasonably be investigated to determine actionable business insights.

The four most important features for the Random Forest Regressor model were *NumProducts, ARP, Market Share Growth (% of Market) Prev. 2 Mo*., and *Average 3 Mo. Units Growth.*

NumProducts, which has consistently proved itself to be an important feature, tops the feature importances for the RFR model, with a relative importance of 7.86x the average importance,

and standard deviation of 0.42x importance. Therefore, the RFR model confirms the takeaways on product variety discussed above for the XGB model.

ARP's importance (2.59x, 0.29x) to the RFR model similarly confirms the retail pricing takeaways mentioned above in the XGB model discussion. Random Forest Regressors fail to indicate a clear directionality of feature effects, just like tree-based extreme gradient boost models, so conclusions about directionality of *ARP* effect remain elusive.

Market Share Growth (% of Market) Prev. 2 Mo. is newly discovered as a potentially important feature (2.25x, 0.39x) in this Random Forest Regressor model. This feature is actually slightly negatively correlated with *Total Sales ($)* (r=-0.088), which may at first seem surprising. However, it is likely that the small handful of massive cannabis companies included in the model had much larger fluctuations in market share growth, since they inhabit a much larger percent of total market sales. Thus, it may actually be true that increased market share growth for small companies increases their next month's revenue, but that noisy dips in market share for large companies clouds this effect when investigating correlations. Future modeling on subsets of the data separating brands based on their typical magnitude of sales might help elucidate the effect of this potentially important feature.

Average 3 Mo. Units Growth was also newly discovered as a potentially important feature (2.22x, 0.46x) for the Random Forest Regressor model. This was uncorrelated with *Total Sales ($)* (r=0.0077) in the original correlation analysis. Unit movement growth over the previous three months seems to be important for predicting total sales in a given month, but the directionality of this effect remains unclear. It is entirely possible that momentum plays a big role in a brand's product sales, and this momentum may come in waves, such that a recent downturn leaves room for increased sales, and prolonged upward trends in unit movement eventually lead to market correction. Cookies should consider conducting a detailed market analysis on their unit movement patterns over time to identify the effects of seasonality, market corrections, and momentum. This may provide insight into expected sales for future months, as well as useful insights which could indicate how to better allocate product for expected unit movement throughout the year.

Data Exploration

Looking at *Total Sales ($)*, there is an extremely long right-skewed distribution. The outliers on sales, generated by a small handful of massive cannabis companies, might be limiting the precision of the ML models analyzed here. While our cross-validation was stratified across total sales quartiles to mediate the effect of outliers being split differently into training/ testing groups, outliers still present a challenge to most ML models. Fortunately, the ensemble models were able to explain variance well despite the outliers.

Average retail price, *ARP,* follows a similar long right-skewed distribution, with more variance among the non-outlier group. The distribution indicates that cannabis brands generally price their average product to sell for less than $100, but some brands extend their average retail

price well above this. Whether these brands are selling larger quantities of cannabis as single units, or pricing according to more premium product lines, this ‘luxury’ business model is much less popular.

Assuming brands all have similar amounts of time series data: less than one third of brands sell ingestibles, most brands sell inhalables, over one third of brands sell concentrates, approximately half of brands sell flower/pre-roll, and about one fifth of brands sell vape cartridges. Cookies can use these distributions to get a better idea of the typical product line composition of their competition.

‘*Months*’, a feature indicating datetimes for each datapoint, was unevenly distributed towards more recent dates, suggesting that either bookkeeping has improved over time, or that more cannabis brands are entering the market. Given the consistent advances in cannabis legalization, the latter is the more likely explanation. Cookies might consider acquiring budding new cannabis companies, or developing a partnership program with smaller companies whereby they discount their products for third-party sale and thus increase their product movement.

Limitations

To perform importance analysis on the XGB and RFR models, a script was developed to perform a KFold Cross-Validation of the model performance while simultaneously tracking feature importances. For tree-based models, such as the *‘gbtree’* XGB model and RFR model included here, the most typical feature importance metric is based on average *gain* generated by a given feature across all trees. Thus, features which give a greater improvement to model performance upon splitting will be granted a larger importance value. While gain is easily interpretable, it should be noted that it is biased to grant higher importance to values which are continuous or high-cardinality. Future research should consider extending this analysis with other importance algorithms, such as feature-shuffling.

Unfortunately, linear regression was a very poor fit for the data at hand. As such, the ensemble models evaluated were given more consideration for causal inference. Ensemble methods are difficult to interpret, as the directionality of feature effects is unclear. Therefore, while the ensemble methods presented here are good statistical models for explaining variance in *Total Sales ($)* and gaining some idea of predictors which require further research, they do not provide context for directional sales effects on their own.

Future Directions

Looking at the two most important binary features from the XGB Model, *Sells Ingestibles* and *Pre-Rolled/Flower*, there is no clear directionality of effect, yet these features were consistently important for XGB model fitting. It is likely that different brands experience different effects depending on their customer base and the brand specialty they may have established. Cookies should perform further analysis on the gross profit margins produced by their edible and flower

product lines to determine whether their brand is benefitting from the inclusion of both kinds of products.

From the data exploration, Cookies can draw several important conclusions about directions for future data analysis.

Looking at the distribution of binary features related to brand inclusion of various product types, there seems to be a gap in the market for vape cartridges. Assuming brands had similar amounts of data on average, only approximately one fifth of the brands included in the data sold vape cartridges at all. Cookies might consider expanding their product line more aggressively into the vape cartridge market to take advantage of this market gap. This business strategy is also supported by the importance of the *Vape Cartridge* feature for the Random Forest Regressor model. As the second highest importance binary feature for the best performing ML model evaluated (RFR), with a consistently moderate relative importance, *Vape Cartridges* might have a significant impact on brand sales.

The models included in this report retain all brands for the sake of gaining a more comprehensive overview of the cannabis market and predictors which influence brand success in a given month. However, the large outliers in *Total Sales ($)* likely reduced model performance. Cookies should consider conducting further research on subsets of the cannabis market based on average brand sales. A good start might be generating separate models for each of the four sales quartiles produced in this analysis: 0-$13,903, 13,903-$62,101, 62,101-$247,327, $247,327+.

Furthermore, given *ARP's* long-tailed right-skewed distribution and consistent importance across the models included in this paper, it would be valuable to extend future analysis to parse brands into groups based on the average retail price of their products, and determine whether more precise models can be fit to the market when stratifying by *ARP.*

6. Conclusion

Using pre-existing cannabis sales datasets, we cleaned, augmented, and imputed them into a single dataset that could be easily used for ML model training and testing. Throughout this process, we removed time series information directly related to raw *Total Sales ($)* values in the hopes of creating a model which would derive more interesting insights. We implemented three statistical regression models, including a basic Linear Regression model, an eXtreme Gradient Boosting (XGB) Regression approach with GridSearchCV-tuned hyperparameters, and a Random Forest Regression (RFR) approach with manual hyperparameter tuning.

Through a 10-fold cross validation, we obtained the following results. Overall performance was the worst for the linear regression model (R^2=0.2639, MAE/μ(sales)=0.9996). Next, the XGB model was great at explaining variance in *Total Sales ($),* but still not particularly precise (R^2=0.8897, MAE/μ(sales)=0.4437). Lastly, the RFR model was the best at explaining variance in *Total Sales ($),* and was the most precise (R^2=0.8961, MAE/μ(sales)=0.3695).

Due to the low model performance for basic linear regression, inferences were made based on the XGB and RFR models, combined with Pearson's correlation coefficients between features and *Total Sales ($)*. Out of all the features, *NumProducts* was the most significant positive predictor, indicating that some combination of product variety and brand size is a powerful indicator of *Total Sales ($).* Since average retail price and recent market share growth were also consistently important predictors, retail pricing and brand momentum are likely also important factors driving monthly sales.

In conclusion, *Total Sales ($)* for future months of cannabis sales can be predicted with moderate accuracy even after removing previous month sales and other features with high collinearity. Upon generating models without these features, numerous avenues for further brand research become obvious. One possible future direction would be to bucket brands based on *Market Share (%)* and generate predictive and inferential models for each bucket. In this way, cannabis brands could be separated based on brand size, and trends could be identified uniquely for new brands, moderate-size brands, and massive cannabis enterprises.