Cookies Business Intelligence Report

Consultants: Noelle Wheeler, Jenny Krystkowiak

Date of Completion: December 4, 2021

I. Executive Summary

Tasked with the goal of applying data science and business intelligence to the globally recognized company Cookies, this report provides data-driven insights to aid in strategic and tactical business decisions relevant to continued company growth and success.

With the goal of forecasting product sales and determining key factors that impact product success, a predictive model was built and optimized to estimate the total sales of a given product with an accuracy of 79.85%.

Feature selection and statistical analysis revealed that key indicators for the likely success of a new product launch include: total THC and CBD levels, item weight, brand, and flavor.

This report seeks to outline and justify the steps taken to arrive at these results, as well as provide a discussion on actionable insights and suggestions for the continuation of data analysis to drive company growth and success.

II. Background/ Introduction

As a top-selling, globally recognized, and still relatively young company, the 2012-founded Cookies has come to dominate the cannabis industry for its diversified and successful product portfolio of smoking and streetwear products. From its origins of a San Francisco garage to global growth, Cookies has utilized pop culture, music industry connections, and innovative genetic engineering to provide high quality and varied products. In addition to offering over 50 cannabis varieties and product lines, Cookies also pioneers a Social Impact Program aimed at targeting issues critical to providing more accessibility and equality in this industry. This program targets these initiatives through public outreach, education, and reinvestment in communities negatively impacted by the War on Drugs.

As a fastly growing and largely spread company, leveraging business intelligence to transform data into actionable insights is an increasingly more relevant goal. As the needs of the company progress into developing plans to provide the same quality of products while lowering cost, increasing manufacturing abilities, and driving the successful advertising and diversification of products are critical components of further company growth. These goals heavily rely on reliable predictions of future sales and insight into the key components of the company's past success in driving higher total sales. This report aims to provide these data-driven insights to aid Cookies in its strategic and tactical business decisions to continue its growth and success.

III. Methodology

In order to extract these business intelligence insights, a predictive model was created to help forecast product sales and feature extraction analysis was leveraged to determine the key factors that likely impact the success of a product.

In order to best make predictions about total sales we choose to look at each individual product for each month. To do this we utilized the "Top50ProductsByTotalSales" dataset and the "Brand Details" dataset. Both dataset contain information on a product level, so we merged these two tables on product. For every row in the "Top50ProductsByTotalSales" table information is included about each product (from "BrandDetails"). This included "Category L1", "Category L2", "Item Weight", and many more details. We did not look at the ARP or Unit data from the Brand Details table because these columns did not include time-specific data. We also dropped "Pax Filter" because the column had only one value in our dataset. As a result, the primary key of our final merged table is Product and Month. Each row in our table should have a different total sales value that corresponds to a certain product in a given month.

Time Series Feature Extraction was first employed to find the rolling average of total monthly sales for each product, which allowed for a linear regression model to be fit to the data and predict total sales for a given product in a given month. The data used for this model was the total sales of the top 50 products, which when merged with the dataset providing details on each product sold, give a dataset that can be used to predict the total sales of a specific product, based on past sales. In order to avoid introducing bias by selecting only features that were perceived to be impactful to product success, only features that had no variability in its input across all products were excluded. Each feature was visually examined using histograms to gain a large-scale awareness of the variability in each feature as well as visually identify the existence of outliers in the data. An analysis of correlation between each feature and the target variable (total sales) was conducted then displayed in a heatmap-type correlation matrix. Categorical variables were OneHotEncoded to convert categorical values to numeric values without ordinality and numerical variables were scaled using StandardScaler, which scales this data to a standard range for better performance when input into the machine learning algorithms used.

After fitting a linear regression model to our data, a random forest ensemble method was employed to generate an optimized prediction model. This method was selected over other ensemble methods as it produced better performance than the gradient boosting ensemble method, which performed worse when assessing model impact on overfitting and variance reduction.

In order to test the effectiveness of the ensemble and single regression training models created, 10-Fold Cross-Validation was employed. This method shuffled the data to randomize the splitting of the data into test and training subsets then ran the models and assessed their accuracy. This process was repeated 10 times, as this was declared a sufficient number to gain a sufficient assessment of the accuracy of these models without compromising computational cost.

IV. Results

We can see that our linear regression model performed much better than our random forest model. The linear regression model had a RMSE of 232,942 whereas the random forest regressor had a much larger RMSE of 1,317,184. This leads to the conclusion that the given data follows a linear distribution, which makes sense because month and year are input variables, suggesting sales are increasing linearly with time. Linear regression also performed better than the Gaussian Process Regressor and the SVM regressor. The SVM regressor is the second best model after linear regression with a RMSE of 742,022. . This implies that the data is linearly separable in some dimension. Surprisingly, when kFold analysis was run with linear regression and random forest, kFold predicted a very high accuracy for Random Forest, and a lower but still high accuracy for linear regression. This leads to the assumption that the random forest is a good predictive model but the simple splitting of the data did not properly train the model. Perhaps the model got a subset of the data that was not a representative sample.

When looking at the calculated p-values, we can see that most variables have a p-value of 0. This means that most input variables in the dataset are important in the prediction of sales. This informs the decision to not to drop any variables. It is difficult to infer importance from the table generated from Ordinary Least Squares because One Hot Encoding the categorical variables created many new input variables and interpreting the meaning of each new p-value is difficult, but not impossible. When looking at the correlation matrix, it is clear certain variables have higher correlation with total sales than others. For example, Generic Vendor and Generic Items are highly correlated with sales. This makes sense because perhaps Generic Vendors generally charge more. Item weight is also highly correlated which makes sense because larger items would typically cost more. Total THC is the third most correlated item with sales. When experimented by taking out those variables, other important predictors were determined due the decreased robustness of our models.

This data analysis revealed that key indicators for the likely success of a new product launch include: total THC and CBD levels, item weight, brand, and flavor.

V. Discussion

When making recommendations to Cookies for selecting new products, the data supports the recommendation of choosing products with higher levels of THC and CBD. Intuitively, the correlation between the total levels of active ingredients in these products and their sales is reasonable, given that it is these active ingredients that can pose the greatest appeal of these products.

The data also suggests that item weight is correlated with total sales, which could be intuitively explained by the expected relationship between amount of product and total price. However, this is still a critical finding in this context, given the significant variability in brand, type, and ratio of active ingredients to inactive ingredients of each product. Due to this reasoning, we do not recommend using item weight as a major factor in choosing how to adjust Cookies' product portfolio.

Flavor and brand also appeared to be strongly correlated with total sales, given that the robustness of the models decreased when these features were removed from the input dataset. Since these features are categorical, further brand-specific analysis would need to be conducted to provide further insight into the top brands that contribute to higher revenue. This analysis could also be supplemented with either feature-engineering that groups all brand-specific flavor descriptions into a set of flavor profiles (ie: tropical, berry, minty) or building brand-specific models. The calculation of linear regression feature importance would be useful in assessing which of these categorical variables are strongly correlated with total sales.

VI. Conclusion

The goal of this business intelligence assessment was to build an accurate predictive model to help forecast product sales and to conduct analysis to determine key factors that could inform Cookies of potential growth areas for its company. Sales and product specific data was collected and prepared for building the machine learning models through scaling, augmentation, imputation, and feature reduction strategies. A single pipeline was created to accomplish this data transformation, which was then used to build a linear regression model that could predict the total estimated sales of a given product. Important features that exhibited a strong correlation to total sales were found by conducting statistical analysis and regression. Feature reduction was then accomplished by implementing a Principal Component Analysis, and a Random Forest model was employed as an optimized prediction model. The training results used in these predictive models were then cross-validated to assess the effectiveness of these models. Finally, parameter tuning was accomplished to further optimize the predictive model.