Group 5
Manpreet Dhindsa (mkd8bb), Nitika Kataria (nk3rf), Olu Omosebi (oo7bq), Mary Youssef (mry8ea)
Final Report Dashboard

What is the business problem or opportunity and why does it matter to the business?

Commonwealth Foods, Inc. (CFI) is a local grocery store founded in Charlottesville, Virginia by several young graduates from the University of Virginia. It is a popular chain due to its healthy food options, environmentally friendly production, and local products. The once quaint local store rapidly grew outside of Virginia, and although their products are fresh and in-demand, they are ultimately similar to other grocery stores, which is why CFI recognizes a need for thoughtful marketing strategies to maximize profits while continuing to please customers.

Their first steps of creating a frequent shopper program and leveraging campaigns aided sales, however the chain failed to complete analysis on their strategies to assess if they were effective. Fortunately, they collected data on a variety of features about the campaigns, coupon redemptions, transactions, demographics, and products that were involved with households and their use of certain coupons with specific stores and products. Dr. Rachel Rosen, the Chief Analytics Officer, has tasked our team to work with Zhora Salome, her fellow Chief Marketing Officer, on evaluating marketing metrics in order to increase their overall revenue.

In the following sections, we will describe our findings from both exploratory data analysis and predictive models. We will be working with Zhora "Z" Salome, the Chief Marketing Officer, to focus on the marketing problems for CFI. We want to focus on the performance of two of CFI's largest marketing investments - promotions (e.g., coupons, etc) and the frequent shopper program and make recommendations based on this. Our team wants to focus on increasing revenue for CFI, so the analysis will focus on that in terms of sales value, quantity of items purchased, as well as overall gross revenue based on certain predictor variables. We also want to focus on profitability, but we need to request more information regarding the cost of the campaigns and marketing investments.

What did you analyze and why? How did you go about it?

Effect of Discounts and Campaigns on Quantity and Sales Value

First, we wanted to learn whether the promotions are working or not. We did this by considering sales value and quantity of items sold as response variables individually with various predictors such as retail discount, coupon discount, coupon match discount, campaign description, and the type of campaign. We decided to first focus on sales value and quantity because those two variables have a direct impact on revenue, which we understand is a key focus to increase for Commonwealth Foods. To further explain, the retail discount refers to the loyalty program, the coupon discount is the discount applied due to manufacturer coupon, the coupon match discount is the discount applied due to CFI's match of manufacturer coupon, the campaign description refers to the type of campaign (A, B, or C), and finally, the type of campaign uniquely identifies the specific campaign (1-30). Additionally, we wanted to understand how to expand the frequent shopper program as it is one of the most important variables to consider for increasing revenue, so we needed to determine what additional investments would be necessary to do so. Finally, we analyzed how marketing techniques affect different stores as well as how marketing techniques and household demographics interact with one another.

For the base model of sales value as the response and quantity, coupon discount, coupon match discount, retail discount, campaign description, and types of campaign as predictors, the best model obtained was Random Forest with a R^2 of 0.495 as shown in Model 1 in the exhibits and the Final Report Dashboard under Models[1]. The most important variables in this model are

[1] *All models and metrics are explained below in Table M.*

the quantity and the retail discount. With this model, it seems that quantity and sales value are highly correlated with each other, so we ran a correlation matrix to confirm this. As shown in Table K and the Final Report Dashboard under Tables, there is a high correlation between quantity and sales value of 0.592, so we chose to remove quantity from the following models because variables that are highly correlated with each other should not be used as response and predictors together to achieve accurate results. This also emphasizes the need to continue the loyalty program as it is clear that retail discount has a strong influence on sales value. However, because types of campaign, coupon discount, description type B, and description type A have very little variable importance, we do need additional information in order to provide strong recommendations on how to expand the campaigns, which will be explained further down.

For the base model of quantity as the response and sales value, coupon discount, coupon match discount, retail discount, campaign description, and types of campaign as predictors, the best model obtained was Random Forest with a R^2 of 0.996 as seen in Model 2. This model performs very well, but it has the same issue as above where sales value and quantity are highly correlated with each other. However, this also shows that retail discount is the most important variable for quantity, which again emphasizes the need to continue the loyalty program. Much less than that are the campaign description of type B, coupon match discounts, campaign description of type A, and coupon match discount. As explained above, we still do need additional information regarding the campaigns, which will be further explained in the recommendation section.

So, next we analyzed sales value as a response against the predictors of retail discount, coupon discount, coupon match discount, campaign description, and the type of campaign. Quantity is removed as a predictor since they are highly correlated with each other. Sales value is an important response variable to consider because this is the total amount of dollars Commonwealth Foods receives from the sale, which directly impacts CFI revenue. By finding the best model based on the highest R^2 score, a Random Forest analysis had the highest performance with a 0.155 R^2 score. While this score is not high enough to be significant, the variable importance analysis still provides valuable information that the retail discount, which is the loyalty program, is the most important variable to this model above all the other variables. This can be seen in Model 4 of the Exhibit and in the Final Report Dashboard under Models.

Other valuable information from this model that should be considered is that the different campaigns and the description types of the different campaigns do not seem to have much of an impact on the sales value in terms of variable importance. So, we believe that Commonwealth Foods should continue the existing campaigns, but we currently do not see the value in expanding on campaigns yet. This is also due to the fact that we lack a lot of information about campaigns such as the cost of each, the target audience for each, the products that the campaigns are focusing on, what the descriptions actually describe, and the discount type that the campaigns are targeting. This will be further discussed in the recommendation section.

Next, because we know that campaigns themselves do not have much impact on sales value, we wanted to narrow down to using only retail discount, coupon discount, and coupon match discount as the predictors when sales value is the response. The best model based on the highest R^2 score is the Random Forest with a R^2 score of 0.178. Despite having an insignificant R^2 value, the model supports the fact that retail discount has a very high variable influence on the sales value compared to the coupon discount and coupon match discount. This can be seen in Exhibit Model 5 and in Final Report Dashboard under Models.

Next, we analyzed quantity as a response against the predictors of retail discount, coupon discount, coupon match discount, campaign description, and the type of campaign. Quantity is important to consider since that represents the number of products purchased during the trip. By finding the best model based on the highest R^2 score, a XGBoost analysis had the highest performance with a 0.514 R^2 score. This score is much higher than seen with these variables against sales value, and the variable importance analysis still provides valuable information that the retail discount, which is the loyalty program, is also the most important variable to this model above all the other variables. This can be seen in Model 3 of the Exhibit and Final Report Dashboard under Models.

This model also confirms what we saw with sales value as a response variable. The different campaigns and the description types of the different campaigns do not seem to have much of an impact on the quantity in terms of variable importance. So, we again believe that Commonwealth Foods should continue the existing campaigns, but we currently do not see the value in expanding on campaigns yet due to the lack of information surrounding campaigns.

Next, we want to narrow down to retail discount, coupon discount, and coupon match discount and exclude information regarding campaigns to see if we can further understand the impact of the individual discounts on quantity as the response. The best model based on the highest R^2 score is the Random Forest with a R^2 score of 0.204. Despite having an insignificant R^2 value, the model supports the fact that retail discount has a very high variable influence on the quantity compared to the coupon discount and coupon match discount. This can be seen in Exhibit 6 and in Final Report Dashboard under Models.

Effects of Discounts on Product Sales and Revenue

Based on the exploratory data analysis of the Commonwealth Foods, Inc's dataset, it was observed that multiple factors, such as product ID, store ID, coupon discount, retail discount, coupon match discount and other variables have an impact on revenue. In order to determine the impact of these variables on revenue, a machine learning approach was implemented with a data set of 20 features from the Campaign, Demographic and Transaction data tables. Four machine learning algorithms namely Random Forest, Ordinary Least Square, XGBoost, and Decision Tree were used to build models to determine the primary features that influence revenue. The results of the machine learning models show that XGboost performs best, with a R-squared value of 0.469, but it favored mostly categorical variables. The performance of the other models was not as good as XGboost with R-squared metrics in the range of 0.323 to 0.415. All the models except OLS identified retail discount as an important feature for predicting revenue. Further investigation into the retail discount shows that there is an influence of retail discounts on the volume of sales at the retail stores. The aggregation of retail discounts and product sales by store location shows a strong relationship that supports the promotion of retail discounts to increase sales revenue.

Improving Sales and Revenue with Demographic Data at Retail Stores

Inspired by a statistician's work at Target who tried to predict if its customers are pregnant, we hope to predict the number of children a shopper has to better focus our campaigns to that specific demographic of parents. The original goal in Target's case was to encourage mothers to shop solely at Target for all those needs. Similarly, we hope to encourage parents to choose CFI as their primary grocery store for all their needs, which would expand on the frequent shopper program. Because the program is designed to reward frequent shoppers with discounts, this would further encourage shoppers to come as the campaigns can target them initially by meeting all their children's needs then rewarding them with discounts, an aspect Target's marketing designed did not account for. By using the customer's estimated age range, marital status, household income, homeowner category, product ID, sales value, and the time of the transaction, we were able to predict the number of children a shopper has with an accuracy of approximately 75% using a random forest model, as seen below in Model 7.

With an accuracy of 75% the model is robust, but it also has an AUC ROC of about 0.93, indicating that the model is not predicting by chance, but is strong in its predictive capability. The strength of the model based on those predictors also indicates the campaigns could be honed to certain customers. This also means that there are shopping patterns based on the above variables, therefore specific recommendations can be made for certain demographics, which will be discussed further below. Surprisingly, when this same model was run with commodity discount rather than product ID, the model seemed to perform worse. We had anticipated that the grouping of products would better group the items individuals with children purchased but it may be too broad of a category thus the products to the detail of their unique ID performed better. A comparison of the two model metrics can be seen in Table L.

What did you learn and what do you recommend to the business?

Recommendation for the Loyalty Program

We recommend that CFI continue to invest in the loyalty program.We learned that through the consumer's use of the retail discount, the loyalty program has the largest impact on sales value, the quantity purchased, and the gross revenue. In addition to the models stated previously, Table I and J in the Exhibit provide secondary evidence to support the influence of the loyalty program through the use of retail discount compared to the coupon discount and coupon match discount. Table I shows the average amount of the three discount types used at the first 25 stores and an 'Others' to aggregate the average discount values from the rest of the stores present in the dataset. It is visible that a large majority of the discount used is the retail discount at all the stores. Table J shows the average amount of the three discount types used for the different products. Even though this table gives insight to the usage of discounts on product level, it still shows retail discount being the majority from the discounts available for consumers to use. Due to the extensive use of retail discounts, the risks to continue the investment into the loyalty program is quite low. Since the program has been running for quite some time, unexpected outcomes are also unlikely. Therefore, we recommend that Commonwealth Food continues to invest in the loyalty program and continue to campaign heavily to emphasize the loyalty program to its current and future consumers to enhance the sales value, quantity purchased, and gross revenue.

Recommendation for the Campaigns

In terms of the campaign descriptions and the types of campaigns though, as stated above, there is too much missing information regarding campaigns, which is why our main recommendation at this point is to continue the existing campaigns but gain more information about them to see how campaigns can potentially be expanded in the future. We cannot recommend expanding campaigns yet due to the lack of information surrounding campaigns. We would need to further understand the cost of each campaign, the target audience for each one, the products that each are intended to support, what the descriptions of campaigns mean, and what discount type is emphasized by each campaign whether that is the loyalty program, the coupon discount, or the coupon match discount. Once we are able to further understand the purpose of each campaign, this will help us create better models that we can use along with the models based on demographics, so that we can target individuals appropriately based on their purchasing behavior. Because campaigns typically have a high cost, there is a risk that the cost will exceed the benefit, so it is important that we gather more information in order to make a decision.

Recommendation for Demographics

To better target specific customer groups, we recommend further data collection and alternative ways of grouping the data to better understand the data and better train the algorithms. To expand on this idea, collecting data on children's gender and age range could help target ads even further for certain toys or gadgets and infant foods or children's snacks. Additionally, the way the products are grouped together under the variable commodity description does not seem to be optimal. When the model was tweaked by predicting with commodity description rather than product ID, the metrics were significantly lower. With an additional meeting with Ms. Salome, a better data dictionary of products could be created and there could be more meaningful definitions with each group of the products. A risk with further data collection is data privacy of customers. Not all customers may be comfortable sharing additional information, especially about their children. Some are very protective of their children so sharing details about their children's ages, genders, and hobbies may be ill-received. Data scientists and the marketing executives must balance customer satisfaction and respecting their privacy and improving their models.

Recommendation for Product Sales and Revenue

One recommendation to the CMO would be to implement a store wide retail discount campaign that would apply to the same products in all store locations. Sometimes discounts are applied to a product in one store whereas there is no discount to the same product at another store location. Implementing the same discount policy to the same products across all stores would increase product sales and improve revenue, and this can be achieved with an online discount program which allows product coupons to be downloaded to electronic devices such as smartphones and used at store locations.

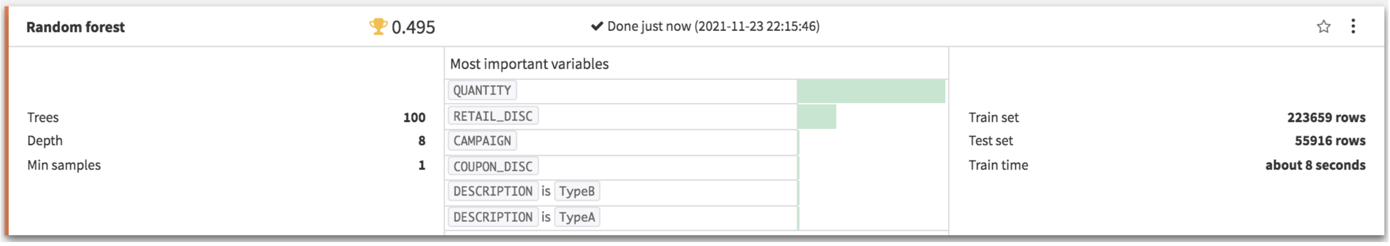
Age group was also identified as an important predictor by the Ordinary Least Square model. Further analysis based on three randomly selected stores, 372, 32004, and 356, showed that each store location had a dominant age group that contributed the highest portion of the revenue. Identifying the dominant age group at each store location, and grouping similar locations together based on age group or other applicable demographic information would make it possible to structure targeted marketing campaigns that would be impactful at the different locations and generate higher revenue.

Recommendation for a Vision Board

As seen with another company that manufactures and sells a variety of consumer packaged goods around the world, we recommend that CFI develops a vision board, so that each store manager can really understand what works for that particular store in terms of market, financial, and operational performance especially since CFI allows store managers to to run their own stores uniquely. By allowing for this much quantitative and data-based decision making, this will enable the individual store managers to take more ownership over the marketing based on analysis of what works based on the data. For example, having knowledge on what our market share is, what our brand demand is, what media is best consumed by our customers such as digital, TV, print, or radio, what our trade channel demand is, what the price point demand is, and even a social media sentiment analysis would be highly beneficial for CFI. By having a vision board that displays this information year over year and allows us to compare directly against competitors, this will allow us to understand where our marketing techniques need to focus in order to increase revenue and reduce cost.

Exhibits

Model 1: Sales_Value as the response and Quantity, Coupon_Disc, Coupon_Match_Disc, Retail Discount, Description, and Campaign as predictors.



Random forest 🏆 0.495 ✓ Done just now (2021-11-23 22:15:46)

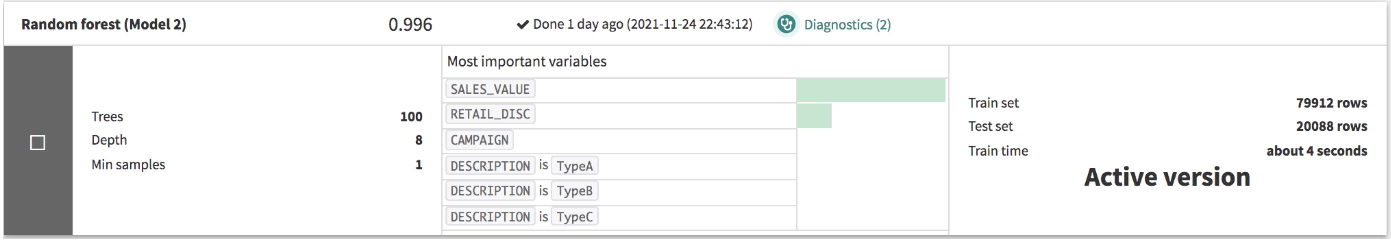
Trees	100
Depth	8
Min samples	1

Most important variables

- QUANTITY
- RETAIL_DISC
- CAMPAIGN
- COUPON_DISC
- DESCRIPTION is TypeB
- DESCRIPTION is TypeA

Train set	223659 rows
Test set	55916 rows
Train time	about 8 seconds

Model 2: Quantity as the response and Sales_Value, Coupon_Disc, Coupon_Match_Disc, Retail Discount, Description, and Campaign as predictors.



Random forest (Model 2) 0.996 ✓ Done 1 day ago (2021-11-24 22:43:12) Diagnostics (2)

□

Trees	100
Depth	8
Min samples	1

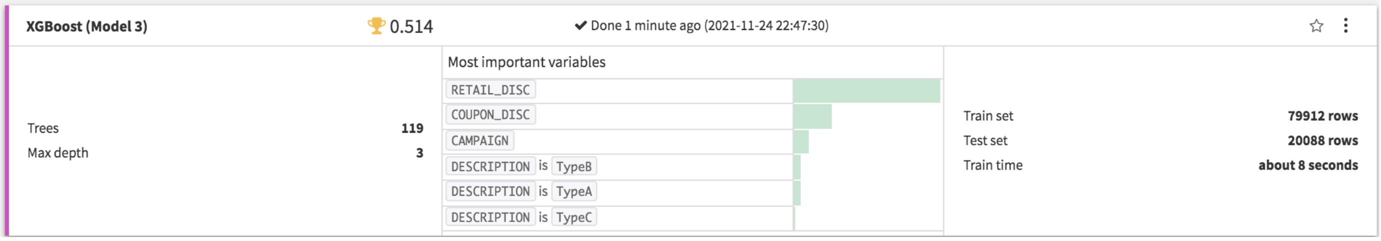
Most important variables

- SALES_VALUE
- RETAIL_DISC
- CAMPAIGN
- DESCRIPTION is TypeA
- DESCRIPTION is TypeB
- DESCRIPTION is TypeC

Train set	79912 rows
Test set	20088 rows
Train time	about 4 seconds

Active version

Model 3: Best regression model based on R2 score with Quantity as the response and Retail_Disc, Coupon_Disc, Coupon_Match_Disc, Description, and Campaign as predictors.



XGBoost (Model 3) 🏆 0.514 ✓ Done 1 minute ago (2021-11-24 22:47:30)

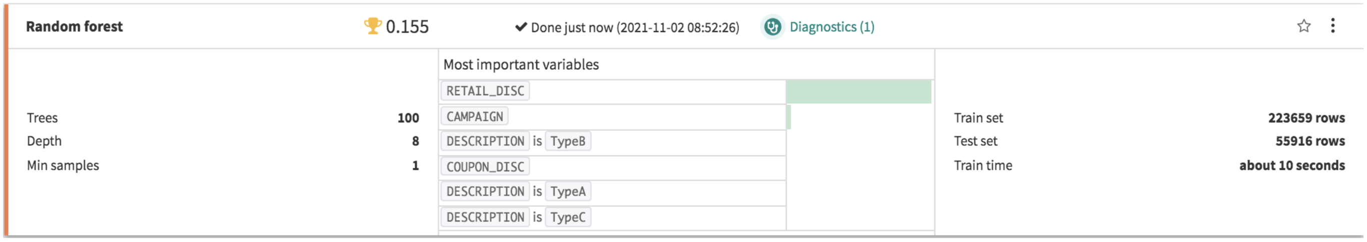
Trees	119
Max depth	3

Most important variables

- RETAIL_DISC
- COUPON_DISC
- CAMPAIGN
- DESCRIPTION is TypeB
- DESCRIPTION is TypeA
- DESCRIPTION is TypeC

Train set	79912 rows
Test set	20088 rows
Train time	about 8 seconds

Model 4: Best regression model based on R2 score with Sales_Value as the response and Retail_Disc, Coupon_Disc, Coupon_Match_Disc, Description, and Campaign as predictors.



Random forest 🏆 0.155 ✓ Done just now (2021-11-02 08:52:26) Diagnostics (1)

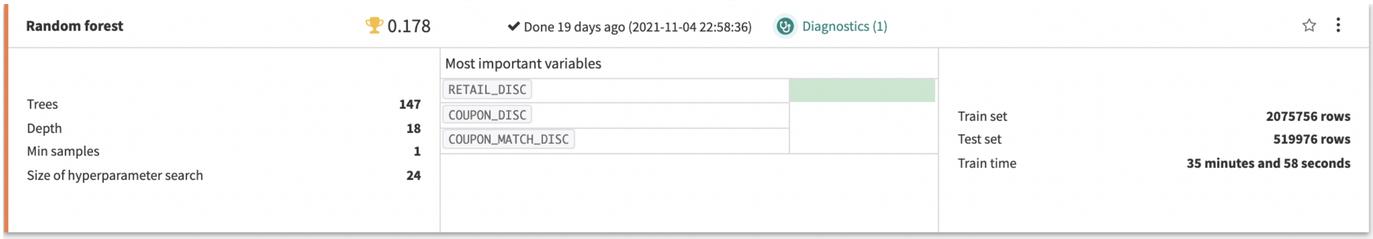
Trees	100
Depth	8
Min samples	1

Most important variables

- RETAIL_DISC
- CAMPAIGN
- DESCRIPTION is TypeB
- COUPON_DISC
- DESCRIPTION is TypeA
- DESCRIPTION is TypeC

Train set	223659 rows
Test set	55916 rows
Train time	about 10 seconds

Model 5: Best regression model based on R2 score with Sales_Value as the response and Retail_Disc, Coupon_Disc, and Coupon_Match_Disc as predictors.



Random forest 🏆 0.178 ✓ Done 19 days ago (2021-11-04 22:58:36) Diagnostics (1)

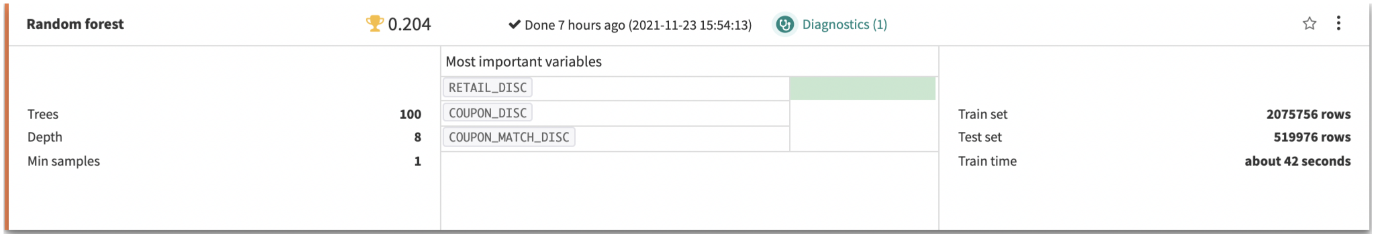
Trees	147
Depth	18
Min samples	1
Size of hyperparameter search	24

Most important variables

- RETAIL_DISC
- COUPON_DISC
- COUPON_MATCH_DISC

Train set	2075756 rows
Test set	519976 rows
Train time	35 minutes and 58 seconds

Model 6: Best regression model based on R2 score with Quantity as the response and Retail_Disc, Coupon_Disc, and Coupon_Match_Disc as predictors.



Random forest 🏆 0.204 ✓ Done 7 hours ago (2021-11-23 15:54:13) Diagnostics (1)

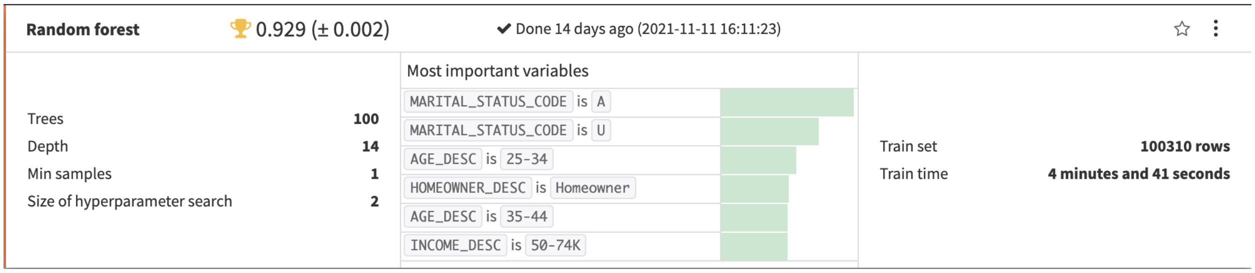
Trees	100
Depth	8
Min samples	1

Most important variables

- RETAIL_DISC
- COUPON_DISC
- COUPON_MATCH_DISC

Train set	2075756 rows
Test set	519976 rows
Train time	about 42 seconds

Model 7: Best random forest model based on AUC ROC for predicting number of children as the response and homeowner_desc, product_id, age_desc, income_desc, marital_status_code, sales_value, campaign, and trans_time as predictors.



Random forest 🏆 0.929 (± 0.002) ✓ Done 14 days ago (2021-11-11 16:11:23)

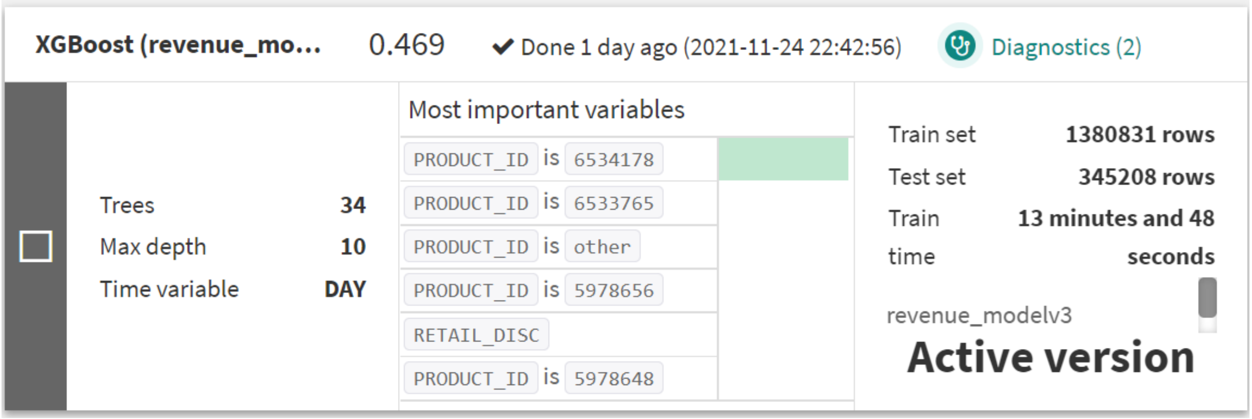
Trees	100
Depth	14
Min samples	1
Size of hyperparameter search	2

Most important variables

- MARITAL_STATUS_CODE is A
- MARITAL_STATUS_CODE is U
- AGE_DESC is 25-34
- HOMEOWNER_DESC is Homeowner
- AGE_DESC is 35-44
- INCOME_DESC is 50-74K

Train set	100310 rows
Train time	4 minutes and 41 seconds

Model 8: XGBoost model to determine primary features that influence revenue. Different models identified other features but XGBoost scored the highest R-squared value.



XGBoost (revenue_mo… 0.469 ✓ Done 1 day ago (2021-11-24 22:42:56) Diagnostics (2)

□

Trees	34
Max depth	10
Time variable	DAY

Most important variables

- PRODUCT_ID is 6534178
- PRODUCT_ID is 6533765
- PRODUCT_ID is other
- PRODUCT_ID is 5978656
- RETAIL_DISC
- PRODUCT_ID is 5978648

Train set	1380831 rows
Test set	345208 rows
Train time	13 minutes and 48 seconds

revenue_modelv3

Active version

Table A: The sum of the quantity based on the different types of descriptions.

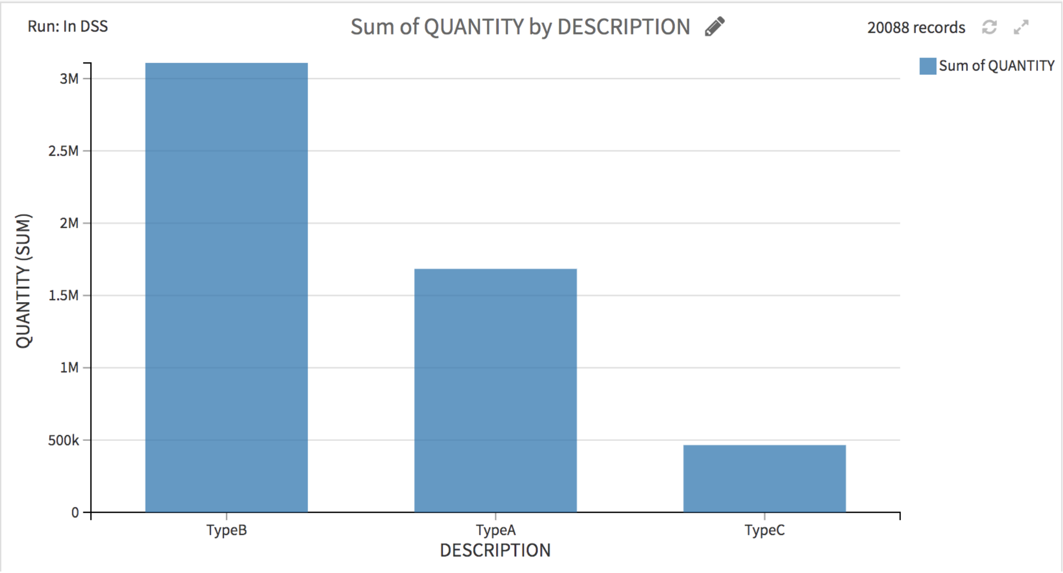


Table B: The sum of the quantity based on the different types of campaigns.

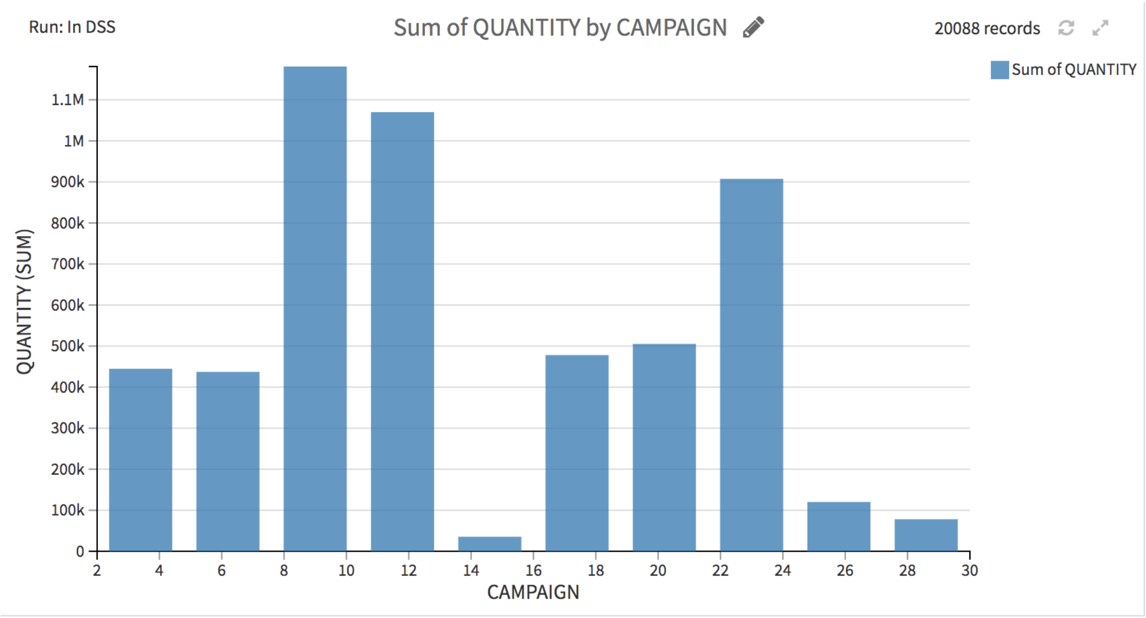


Table C: The sum of the sales value based on the different types of descriptions.

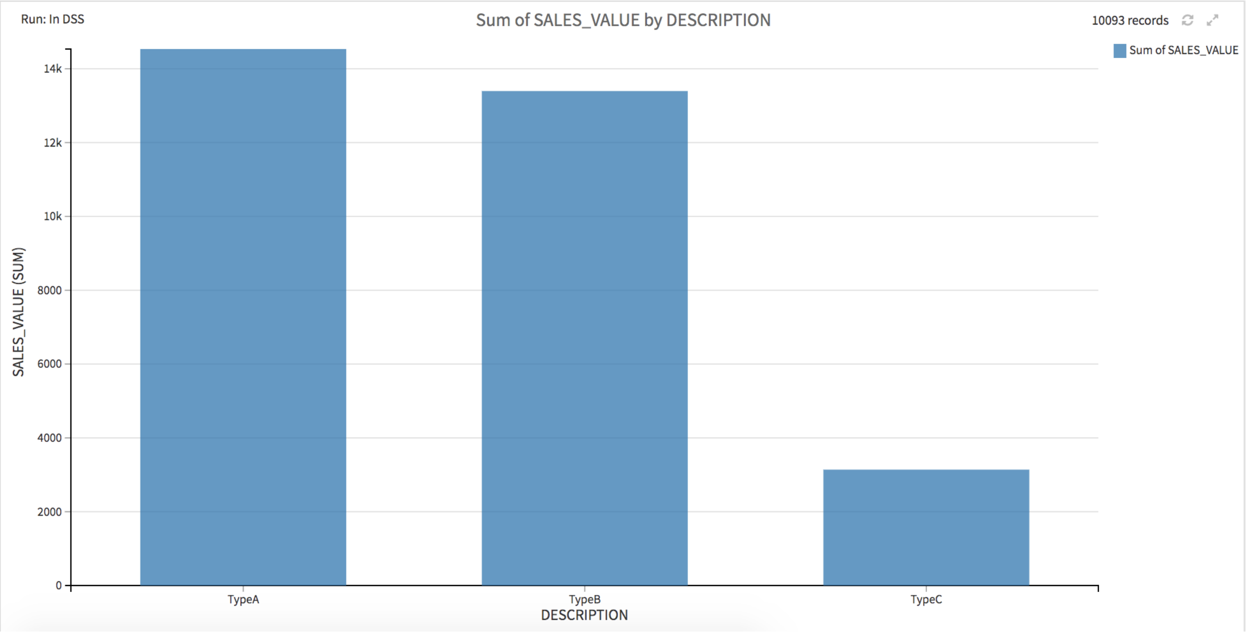


Table D: The sum of the sales value based on the different types of campaigns.

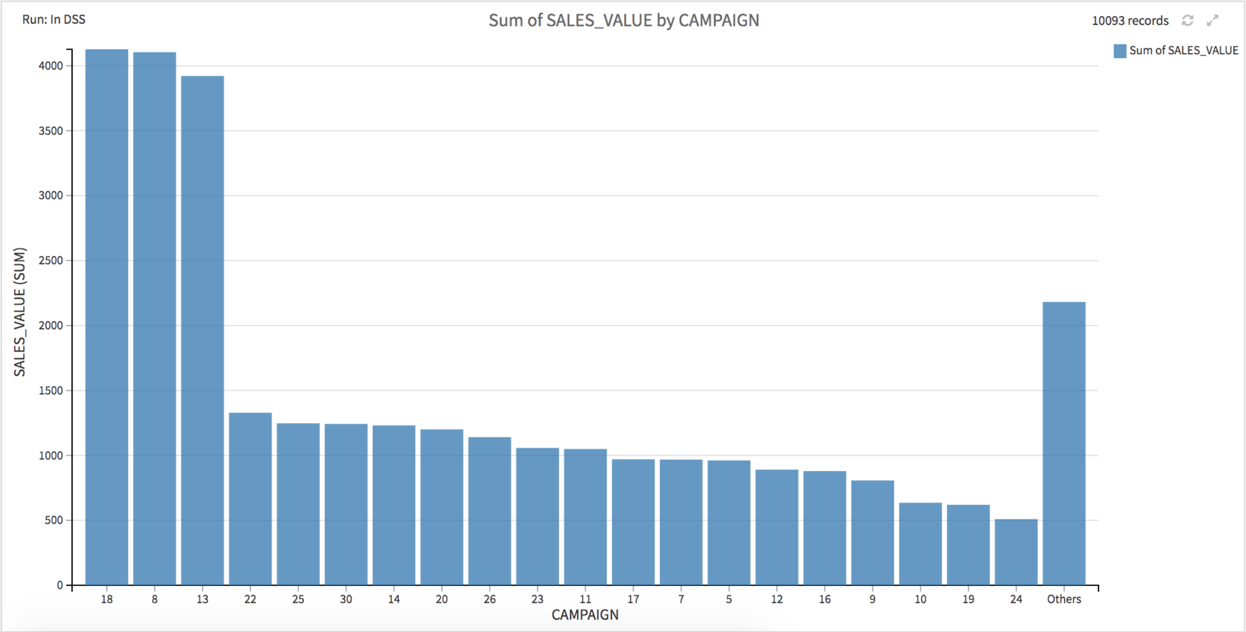


Table E: The sum of the coupon match discount based on the different types of descriptions.

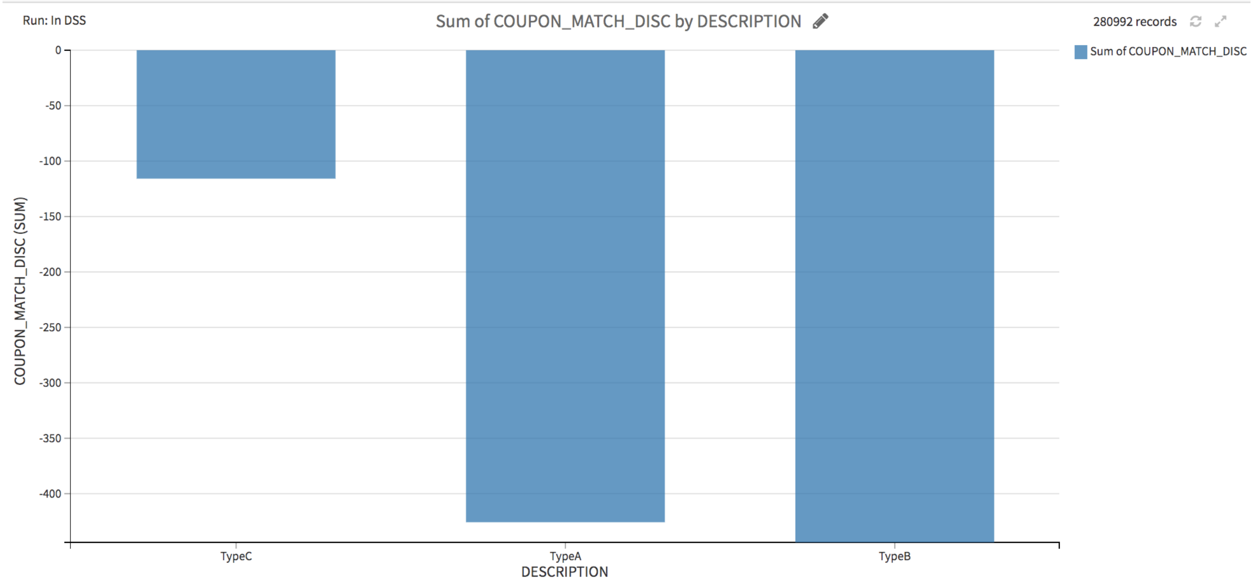


Table F: The sum of the coupon match discount based on the different types of campaigns.

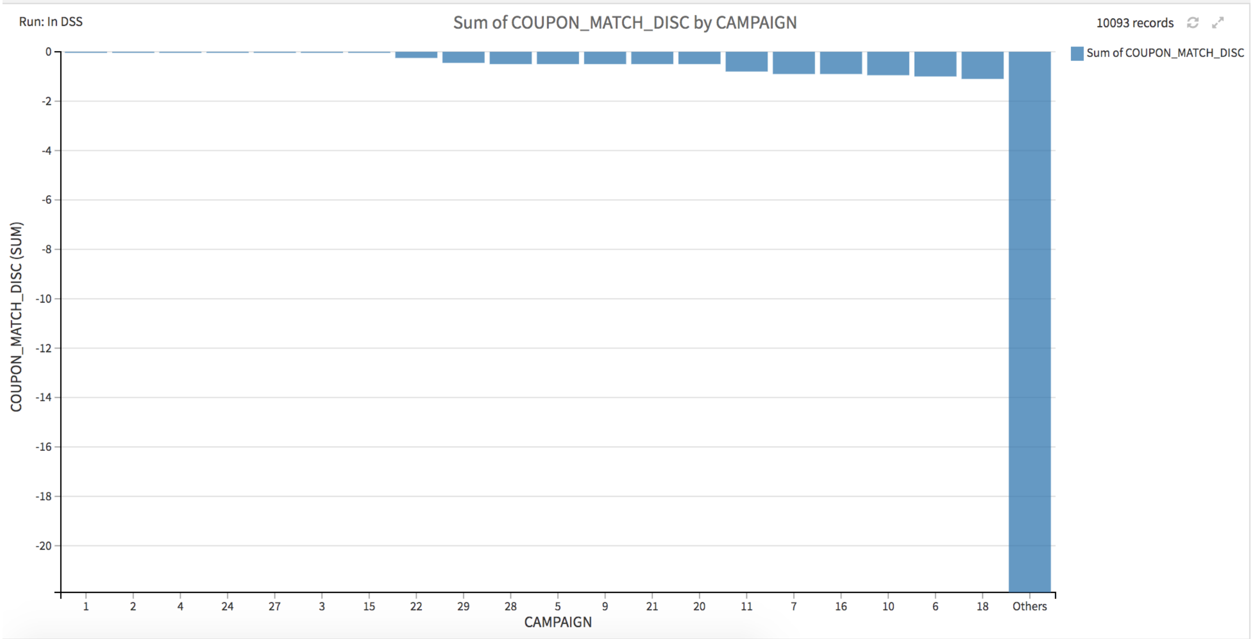


Table G: The sum of the coupon discount based on the different types of descriptions.

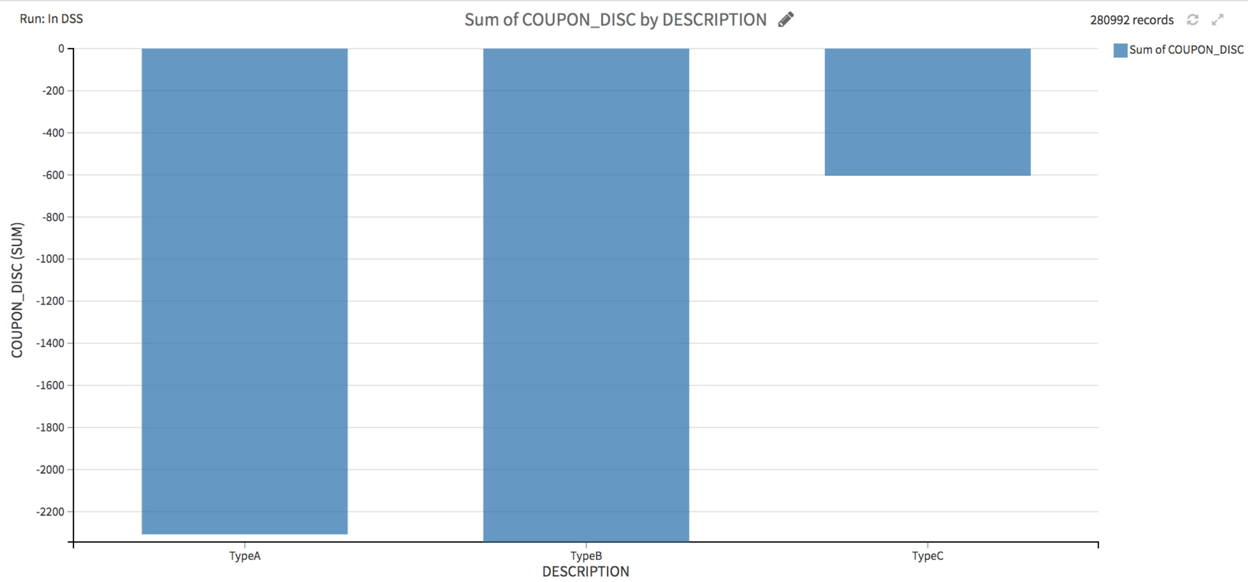


Table H: The sum of the coupon discount based on the different types of campaigns.

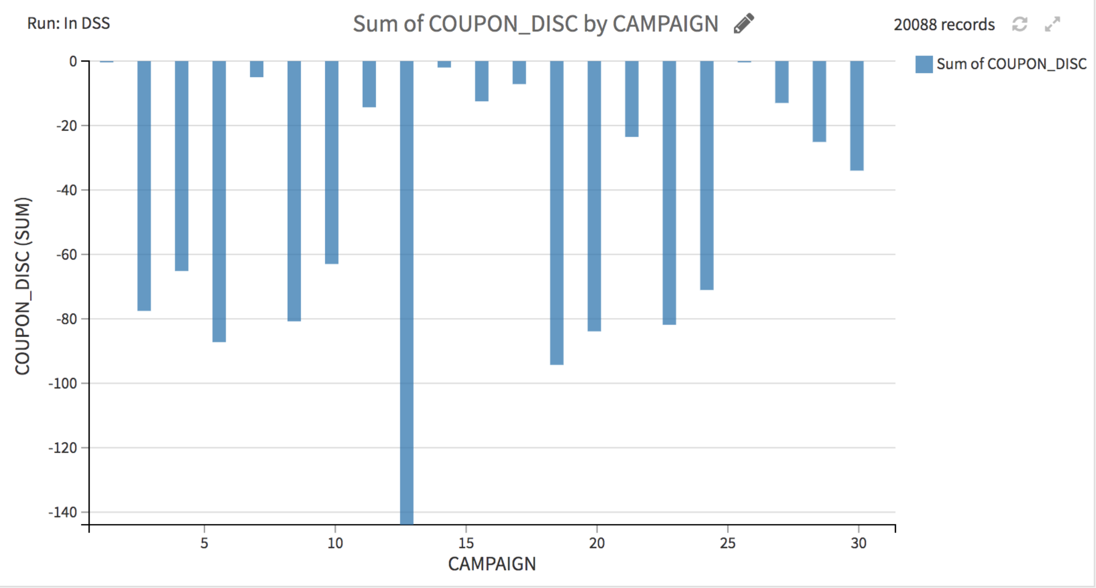


Table I: The average amount of the three discount types used at the first 25 stores.

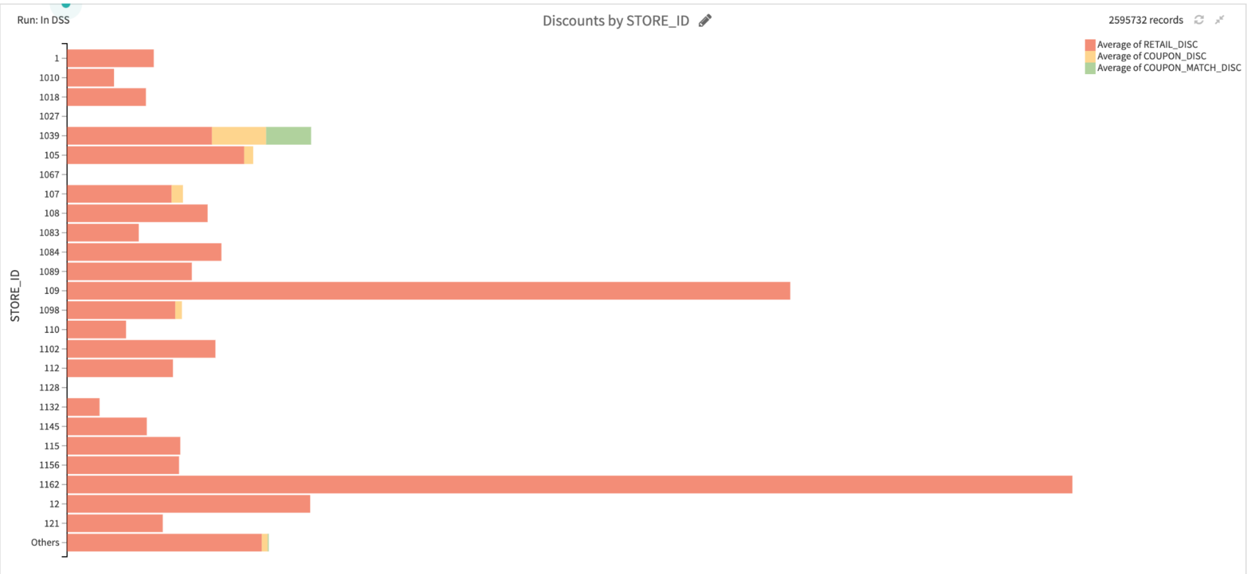


Table J: The average amount of the three discount types used for the different products.

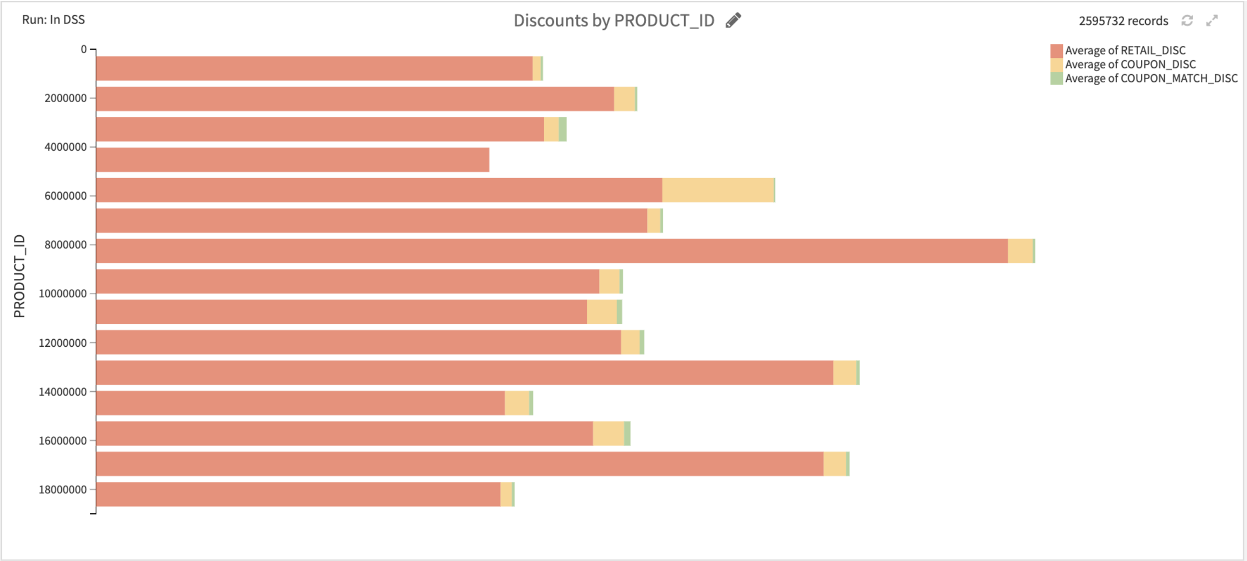
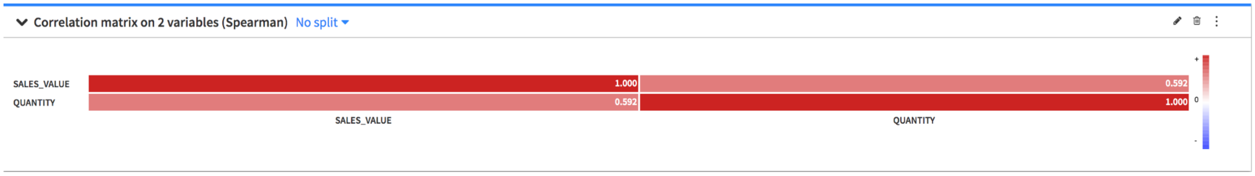


Table K: Correlation matrix between quantity and sales value.


Correlation matrix on 2 variables (Spearman) No split

	SALES_VALUE	QUANTITY
SALES_VALUE	1.000	0.592
QUANTITY	0.592	1.000

Table L: Meaning of Models and Metrics

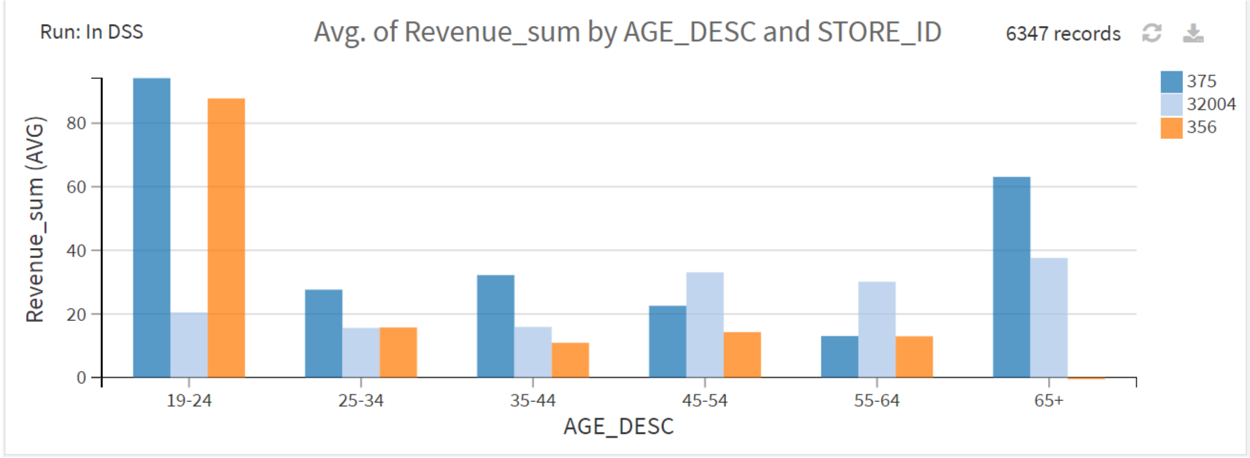
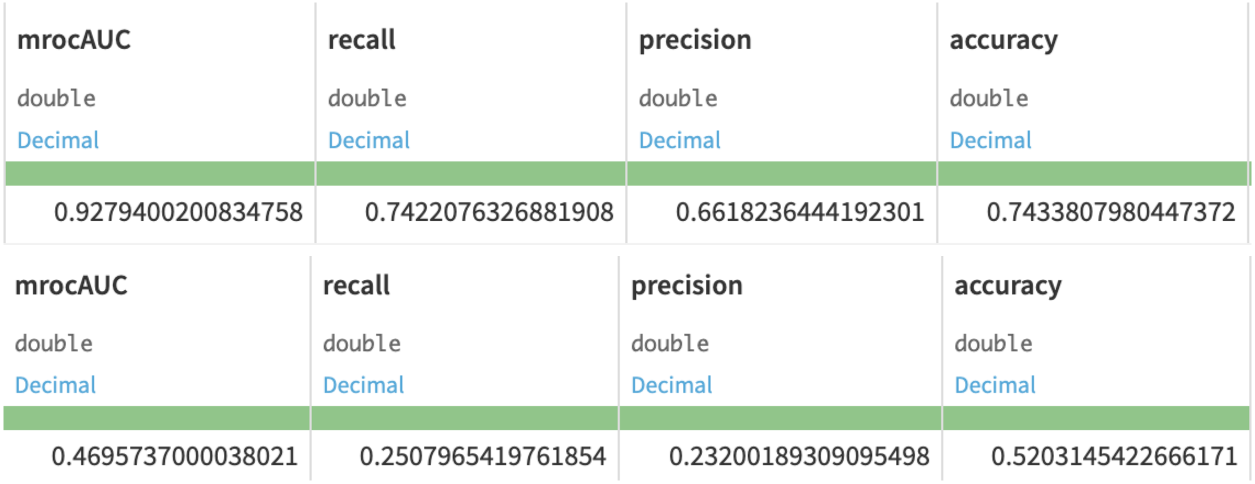


Table M: Comparison of model using Product ID vs. Commodity Description.



mrocAUC	recall	precision	accuracy
double	double	double	double
Decimal	Decimal	Decimal	Decimal
0.9279400200834758	0.7422076326881908	0.6618236444192301	0.7433807980447372

mrocAUC	recall	precision	accuracy
double	double	double	double
Decimal	Decimal	Decimal	Decimal
0.4695737000038021	0.2507965419761854	0.23200189309095498	0.5203145422666171

Table N: Meaning of Models and Metrics

Metrics/Model	**Definition**
AUC ROC	ROC AUC stands for the Receiver Operating Characteristics and Area Under the Curve. It is a performance measurement for classification problems with various threshold settings. The ROC is the probability curve while the AUC is the area beneath it. It's typically best for the graph to be near the top left corner indicating a higher probability. The higher the AUC, the better the model is at classifying between variables.
R^2	A statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared explains to what extent the variance of one variable explains the variance of the second variable. Typically, the higher the R^2 value, the better the model's performance.

Random Forest	A predictive model that uses many individual decision trees that work together as an ensemble and ‘vote’ on the top prediction
XGBoost	XGBoost stands for eXtreme Gradient Boosting and is an algorithm that is an implementation of gradient boosted decision trees designed for speed and performance.