ISM 6136 Data Mining Project

On Predicting

Online Shoppers Purchasing Intention

Created by:

Aditi Kochar
Jesid Acosta
Pranali Kanade
Vidhi Jadav

Muma College of Business

University of South Florida

Table of Contents

1. Introduction

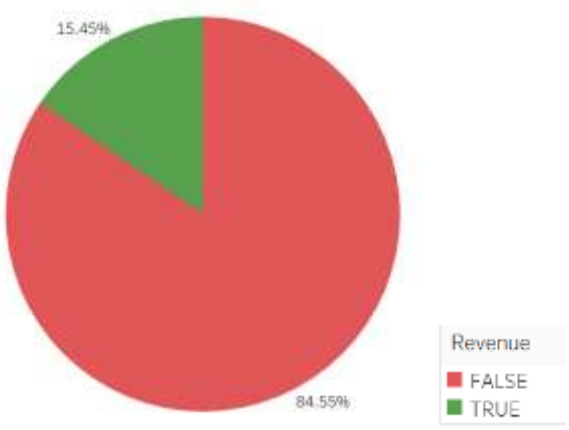
In offline retail stores, the salesperson offers products based on customer requirements with personal attention. Further, shopping moved internet based and a myriad of websites came up with online marketplace. The advent of e-commerce gave rise to offering services with simple clicks from the entire catalogue, customizing recommendations around customers. We have come up with predicting customers who are likely to turn the click through rates into conversion rates by leveraging customer behaviour analysis model built in Azure ML.

2. Dataset Overview

The inspiration and source of dataset is picked from UCI website where aggregate transaction data is collected from a Columbian retail website called Columbia.com.tr outdoors Apparel Company. It is comprised of 18 attributes and 12310 records with Revenue as the target variable. Revenue is a categorical label with Boolean value, TRUE means that the person purchased a product from the website and FALSE means the person did not purchase any product.

A significant independent variable we considered is Visitor Type, being categorical it has 3 values- Returning Visitor, New Visitor and Others. Since Others accounted for only 85 records, we excluded those. We were able to achieve better accuracy out of 12225 records after removing 85 records.

- Total 12,225 records (without 'Other' Visitor Type)
    - 17 Attributes
    - 1 Target Variable "Revenue"
        - 10,551 Negative (Did not purchase)
        - 1,694 Positive (Purchased)



Online Purchase Pie

3. Problem Statement

Revenue is immensely dependent on customer behaviour. Using predictive models, online stores can make real-time behavior analyses and provide purchase incentives for shoppers. The objective of our project is to build a model to improve prediction of revenue generated by applying different models depending on type of visitors. Our further approach is to increase purchases by conducting A/B testing.

4. Exploratory Data Analysis

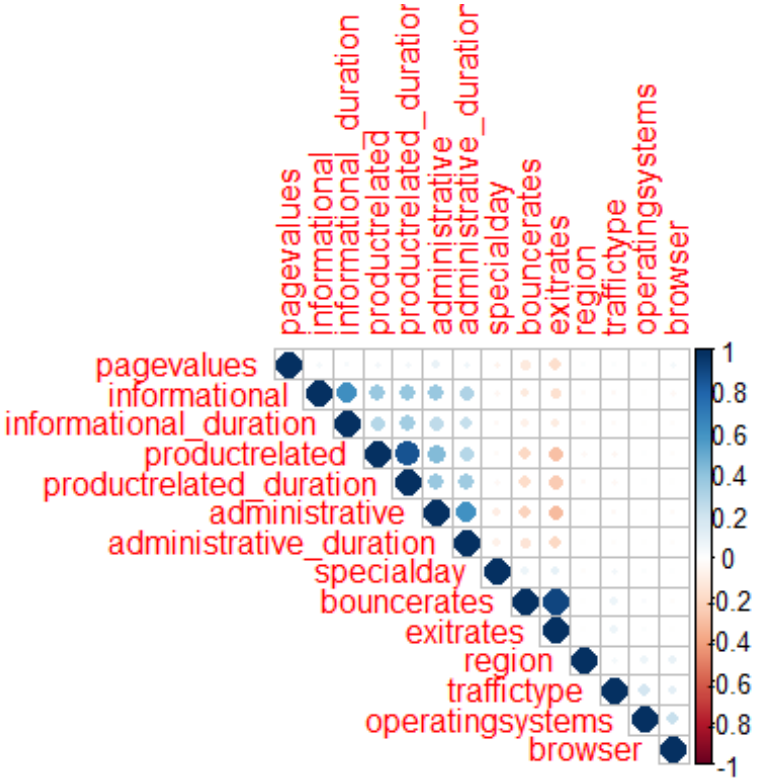
We availed Tableau story board and R programming language for data visualization.

4.1. Correlation between Features

Correlation analysis is a statistical evaluation method to identify the relation between variables. We used R programming language to find interdependency between independent features.

R Code:

```
#Correlation Plot
my_cor = ol_shop[,c(-11,-16,-17,-18)]
correlations = cor(my_cor)
corrplot(correlations, type='upper', order='hclust',number.cex=.5, )
```

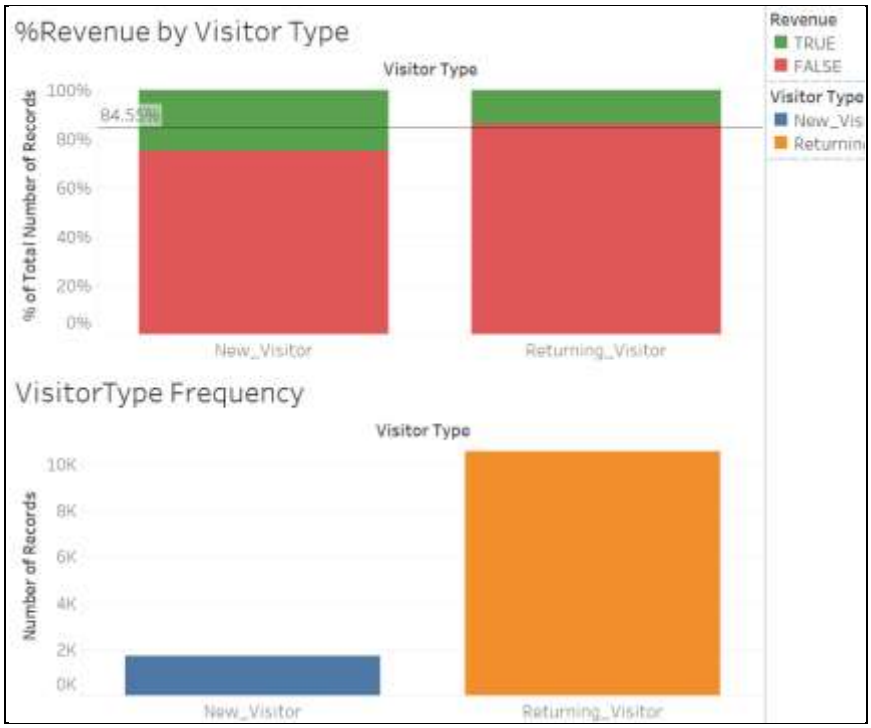


Co-relation Plot

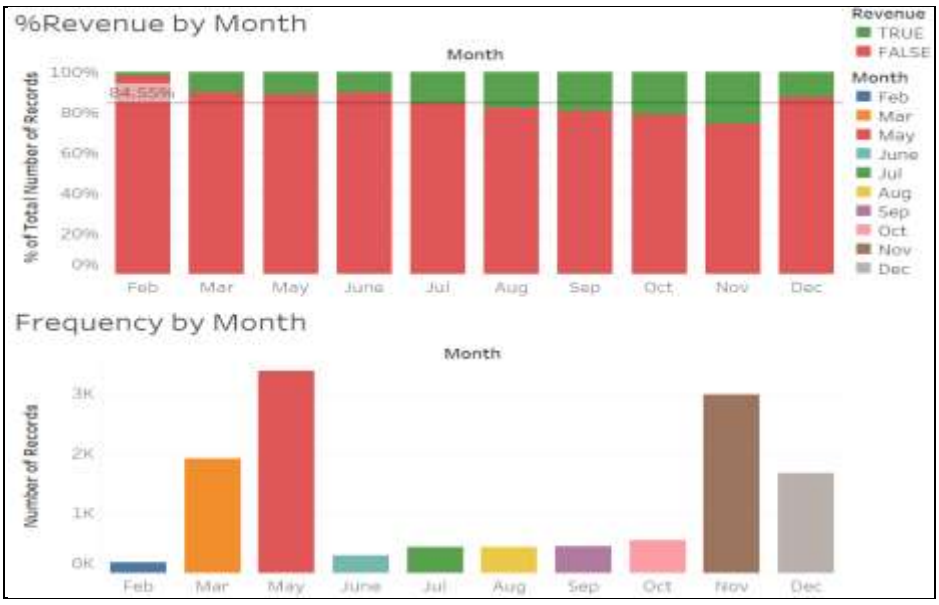
From the above plot it is evident that there are two high correlations (90%) amongst all the independent variables.

1. Bouncerates & Exitrates
2. Productrelated & Productrelated Duration
3. Others either have weak correlation or they are negatively correlated

4.2. Purchase outcome by visitor type



4.3. Purchase Outcome by Month



5. Data Modelling

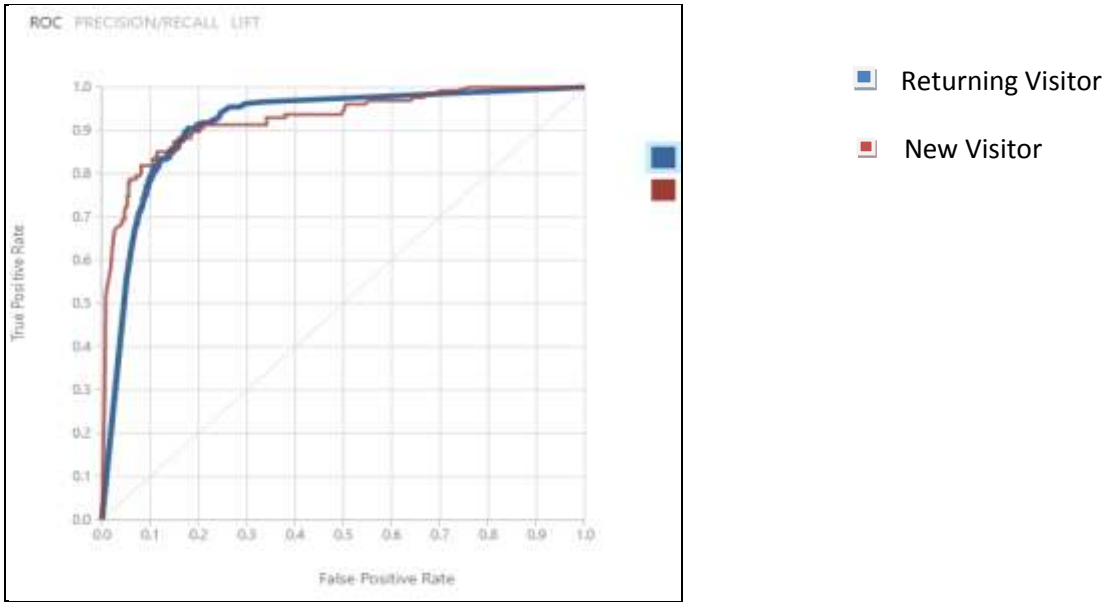
We carried out our analysis by looking at four different models with and without under-sampling.

- Boosted Decision Tree (BDT)
- Locally-Deep SVM (LD SVM)
- Bayes Point Machine
- Neural Network (NN)

Model	Accuracy	Precission	Recall	F1
BDT Base Undersampling	0.844	0.498	0.845	0.627
BDT Base	0.892	0.651	0.655	0.653
LD SVM Undersampling	0.858	0.527	0.769	0.626
LD SVM	0.882	0.745	0.361	0.486
Bayes Point Undersampling	0.851	0.512	0.708	0.594
Bayes Point	0.88	0.781	0.308	0.442
NN Undersampling	0.17	0.157	1	0.271
NN	0.894	0.745	0.474	0.579

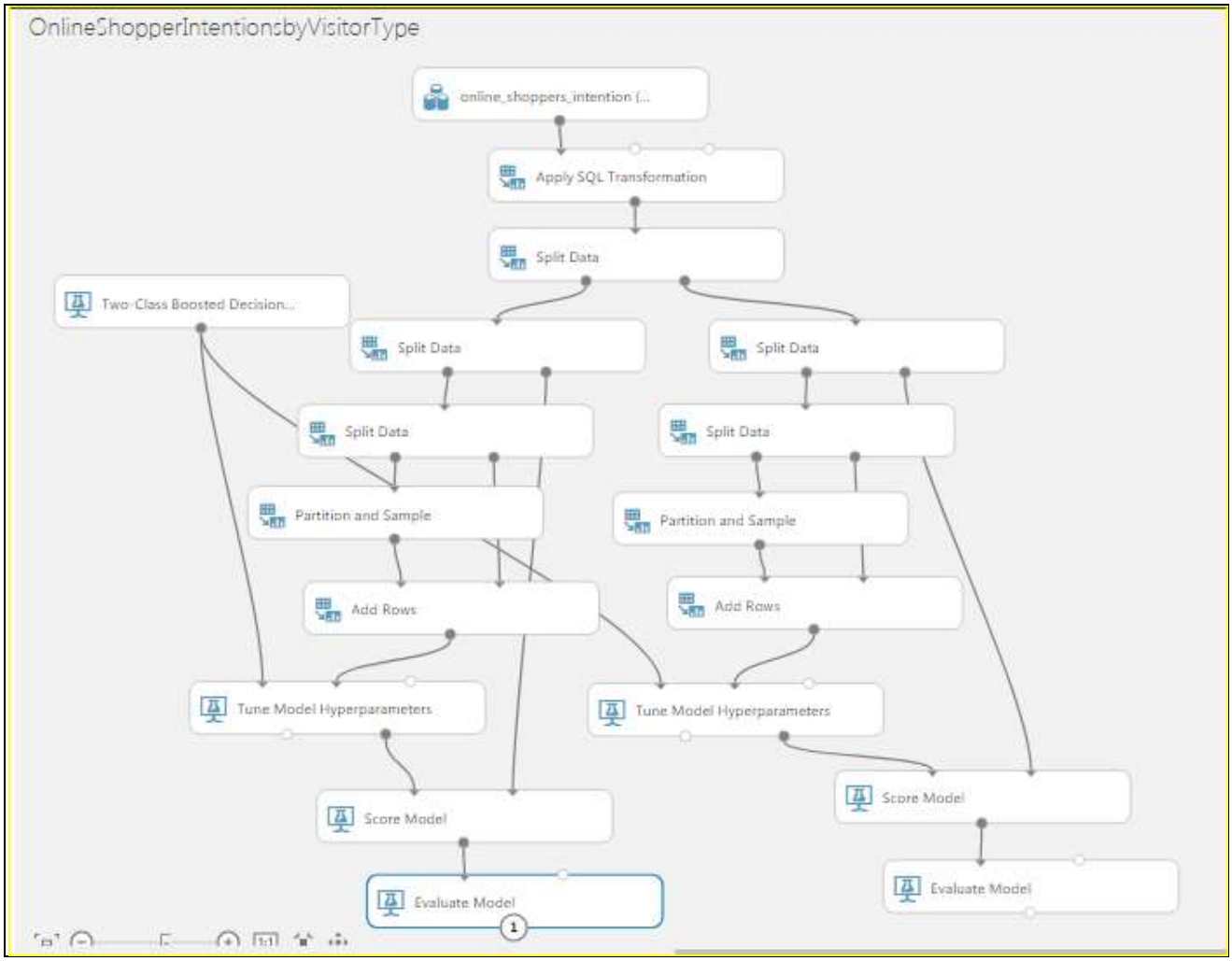
Two-Class Boosted Decision Tree gave us the best outcome.

Model	TP	FN	FP	TN	Accuracy	Precission	Recall	F1
BDT Returning Customers	368	73	358	2366	0.864	0.507	0.834	0.631
BDT New Visitor	115	12	79	303	0.821	0.593	0.906	0.717
Total & Weighted-Average	**483**	**85**	**437**	**2669**	**0.858**	**0.519**	**0.844**	**0.643**
BDT Undersampling	480	88	484	2622	0.844	0.498	0.845	0.627
BDT	372	196	199	2907	0.892	0.651	0.655	0.653



Cumulative Lift

6. Predictive Model in AzureML



Experiment in Microsoft Azure ML

Explanation

Apply SQL Transformation –

After Importing the data, we removed the visitor type – Others.

Split Data –

- In the first split component we split the returning and new-visitors.
- In the second one, we segregated train and test data into 70-30.
- The third split was used to split the revenue based on True (1) & False (0).

Partition and Sample –

Here we under-sampled the False revenue as they were around 90% more in number to 20%.

Add rows –

The rows with True revenue were further added.

Tune Model Hyperparameters –

Here we used the integrated tune and support method to set various parameters like sweeping mode (Random sweep selected), target column, metric for measuring performance (was set as recall), etc.

As mentioned above, Two-ClassBoosted Decision Tree was used and further the model was scored and evaluated after running the model.

7. Observations, Roadblocks and Suggested Solutions

1. More people visiting the pages but lesser buyers
    - ✓ Conduct a survey to understand what is wrong – website layout, products, costs, delivery issues etc
    - ✓ Provide customized discount coupons and offers on products
2. Lessened purchases from returning visitors than new visitors.
    - ✓ Give offers on next purchases
    - ✓ Offer customized birthday coupons
3. Increased sales in the months - May have more visitors
    - ✓ Customize offers based on the special days in May. Eg. May 2 is National Baby Day, we can provide buy 1 get 1 offers on kids clothes/ accessories
4. Lowered sales in the month of February
    - ✓ Increase offers during and around Valentine's Day

8. Summary

The model showed how we can increase predictive performance by using under sampling and a differentiated model for returning visitors vs. new visitors. We found insights from the data that could help us further tailor our website to increase purchasing visitors.

1. We can create slightly better predictions with under sampling and differentiated models by customer type
2. The graphs provide insights on –
    - Amount of pages visited by type of customers
    - Revenue generated
    - Sales on monthly basis
3. Statistics on correlated variables is significant

9. Future Prospects and Recommendations

To test the findings of our data we recommend that two experiments be ran.

In this project we assume that an algorithm (i.e. Two-Class Boosted Decision Tree) will already be used by the online retailer to predict the online-visitor's intention using real-time data by assigning a score in a range from 0 to 1, with 1 being the highest likelihood of purchase. As the retailer gathers this information in real-time from visitors it would determine which are likely to purchase and offer incentives like a one-time promotion. This first experiment detailed below, we assume will show favorable sign of using the prediction model along with promotions to increase revenue, by measuring the number of converted buyers from visits. The second experiment will test our enhanced method of under-sampling and using dedicated models for different types of visitors to see if our new method leads to an increase in purchases.

1. Testing the effect of promotions

We would first want to test the effects of the promotion by using an A/B experiment where visitors would be randomly assigned to either a control group or a group that receives promotion. Using the Two-Class BDT model, visitors in both groups would be scored given their browsing behavior. At the end we would compare the number of buyers in both groups. If the number of buyers for the group that received promotions is higher than the number of buyers in the control group, we can determine that promotions are a good tool to convert customers. Also, we should expect to see that those with scores closer to one are more likely to purchase. If we find this trend this would also help re-affirm the predictive outcome of our model.

2. Testing the enhanced model

Assuming that the promotion tool is effective in turning more visitors into buyers we would now use A/B groups to run an experiment to compare our enhanced model (under-sampling & dedicated model by visitor type) versus the base Two-Class BDT model used in the previous experiment. In this case we would offer promotions to both groups and compare the number of buyers in each group. Given that the enhanced model is driven also by type of visitor we would want to also compare these two classes exclusively, so this will be an attribute that will need to be carefully gathered during testing. If more purchases made by promotion-receiving visitors identified as likely buyers in the second model, we can verify that our enhanced model is allowing us to identify more potential buyers than the base model.

Aside from the experiments mentioned above, we also recommend that the online retailer run trials and multivariate analysis on these following areas that can also help increase sales.

1. To generate more revenue
    - ✓ Partnering with deal makers like Groupon, CouponX, etc to get customers into the doors and test the response
    - ✓ Sponsoring a coupon offer to trigger the customer behaviour

2. Combining these as a combination of 4 scenarios
    - ✓ Riding on advertising tool for the marketing plan to showcase the deals for visitors and how well the company is doing in retaining the loyal customers
    - ✓ Piggybacking on the sustainability goals the company has achieved

3. Conduct Multivariate analysis to predict revenue using previous revenue, conversion rates, page rank, hours spent on the portal, month of the year and visiting customers