Machine Learning with Ecommerce Data

DSC550-T301

Joshua greenert

2022

**Introduction**

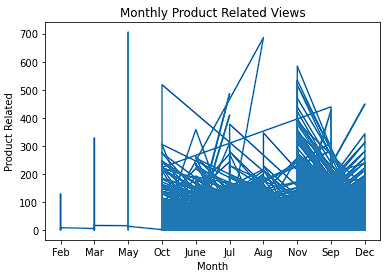
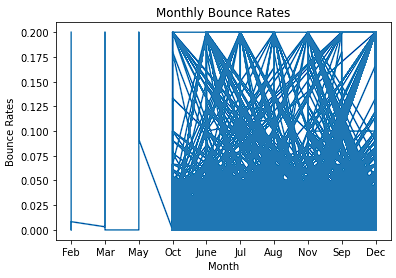
In the world of ecommerce, the flow of data is constant. As smaller businesses enter the atmosphere, they too need to understand the structure of the data and how it can pinpoint areas of opportunity for their respective business. However, without having some guidance on what key indicators predict revenue-generating processes, it can become difficult to survive amongst larger organizations. Moreover, attempts at marketing their business can become wasteful without catering to the right users, holiday events, or distinct expectations.

With indicators specified on how revenue can be generated from the data, users, and businesses alike, can set up their websites to gain the largest amount of revenue possible. Similar to A B testing, these small businesses can understand what is working, and what isn’t working, with their current products and overall design strategies. Solving this problem would not only gain the business a stronger amount of revenue, but a better focus on overall company milestones necessary to support a healthy business.

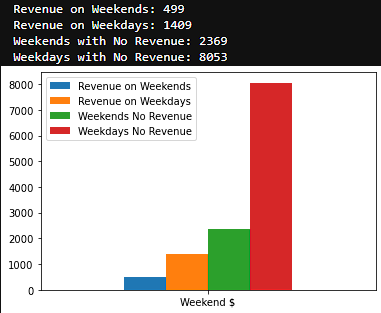
To appropriately pitch this problem to stakeholders, we would need to first focus on the underlying issues within the business. Generally speaking, this means isolating the value that’s being returned in terms of profit for the business as a whole. We can create graphs that express the company's projected growth with a focused marketing plan alongside the actual anticipated growth of the company. With these visuals, our audience can better understand the direction that is expected to be taken to ensure the highest level of success.

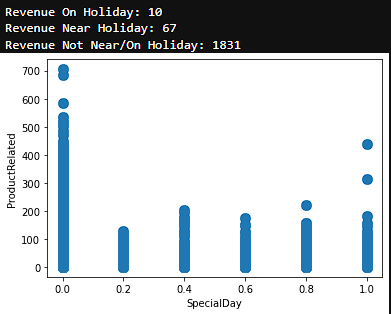
The data for this exploratory data analysis was obtained from Kaggle and includes 12,330 unique user sessions. Additionally, special care was provided by the authors of the dataset to ensure that the data spanned across one year while not being too heavily focused on any particular campaign or special occasion.

**Exploratory Data Analysis**

When initially looking into this dataset, I created some graphs regarding the bounce rate — the number of viewers of a web page that leave after viewing only one page — by month and the product-related views. This overview gave insight that the author may have missed some data when initially putting it all together. Expressed from the graphs, it’s clear that the data for January, February, March, April, and May was missing.

To gain more insight into the revenue column, and when it was taking place, I created another graph to dissect the weekend column based on when revenue was and wasn’t successful (the values were either true or false). The initial idea was to determine if revenue-related purchases were made primarily on the weekends or on the weekdays; the assumption was that weekends would be filled with more conversions than the weekdays as people tend to do their shopping on the weekends. Additionally, the data that expressed what days no revenue took place were informative as well.



Likewise, I performed a similar review on the special day column to determine if revenue estimates were easily visible based on the timeline/proximity to special holiday events. To better visualize this data, I established a basis of collecting data where the revenue event took place on the special, near the special, and not near the special day at all. Therefore, all values were based on when revenue took place.

From the data discovered above, it was easy to determine that some of the initial assumptions were false. If the weekends count for two days, and the weekdays count for five, then the revenue that happened on the weekends accounted for 249 (rounded down) instances per weekend day; whereas conversion that took place on the weekdays accounted for 281 (rounded down) per weekday. Moreover, the information regarding special days/holidays and their ability to increase conversions weren’t statistically significant. The number of days where revenue-related events took place not near or on a holiday dwarfed the amount of instances that were.

**Data Preparation**

The data preparation steps were straightforward. Dummies were created on the categorical columns for the VisitorType and Month columns so that they could be easily reviewed by machine learning steps later in the process. Non-revenue related columns were removed, such as Administrative and Administrative\_Duration, since they didn’t contribute to predicting the target and simply just provided data over admin users, and what they viewed to isolate them from actual consumers. Finally, the fields that held float values were rounded to two decimal places to provide a more visually appealing set of values.

**Model Building and Evaluation**

For the model building and evaluation portion, the data was split into a test and training set with the target being the revenue column. An initial step of hyperparameter tuning was performed to determine which classifier would be preferred over logistic regression and random forest, with random forest being the victor. A second step was then performed to judge between the winner of the last classifier testing with the decision tree, support vector, multinomial Naive Bayes, and gradient booster classifiers; the winner from these classifiers was the gradient booster classifier with a best score of 92.76%.

Next, I reviewed feature selection with chi-squared to select the top three features. The top three best features recommended by this method were Informational\_Duration, ProductRelated\_Duration, and PageValues; this result was a bit unexpected but saved for later use in the evaluation process. Using a decision tree classifier to assist with a predictive model, an accuracy score was found on the training set to produce 99.51% accuracy. Finally, two sets of predictions were created based on the gradient classifier — one with the test features, and one with the recommended features — and the R2 (Coefficient of Determination), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error) metrics were created from each respective dataset.

**Conclusion**

The model created at the end of the process gave potentially valuable results. When utilized with the test data aforementioned, it was discovered that the recommended features were statistically significant. When tested against all features, the coefficient of determination approximated 96.04% while the recommended features accounted for 94.71% of that total amount. Essentially, this provides the impression that the model is ready for deployment. However, substantially more testing is required to ensure that the model is functioning as expected. More data should be collected to properly test the expectation of conversions, and to confirm our model’s conclusions. With these final recommendations, it’s likely that there will be additional minor changes required to complete a model that can better predict actual conversions made by consumers.