How Much Will You Spend on Black Friday This Year?

Nora Myer
The Ohio State University
Columbus, USA
myer.41@osu.edu

Haseeb Javed
The Ohio State University
Columbus, USA
javed.19@osu.edu

Aidan Globus
The Ohio State University
Columbus, USA
globus.3@osu.edu

***Abstract*—Black Friday continues to be one of the most profitable days for retailers in the US every year, so who is really going out shopping every year? Using data about consumer demographics and purchase history on Black Friday, we want to build a model to predict how much money a person is going to spend on Black Friday given a set of attributes including their age, gender, occupation, and marital status [1].**

I. Introduction

Predicting consumer spending behavior has been an active research area in the domain of data analytics. Companies and business enterprises spend millions of dollars on analyzing consumer purchase histories in order to obtain insights as to which factors - be it socio-economic, political, cultural, etc. - affect how much consumers spend and on what. Being able to understand which of these factors are significant is crucial for business as it enables them, among other things, to customize their marketing campaigns for fine-grained consumer segments allowing them to maximize the return on investment for each dollar spent on advertisement. Targeted advertisement campaigns have become even more so prevalent in the era of social networks. Retailers obtain data consumer data from these platforms, use various models to analyze the data and then come up with ads tailored made for individual customers.

Black Friday is one of the biggest shopping holidays in the US. Various retailers, traditional as well as online, offer cut-throat discount offers encouraging consumers to spend a significant portion of their annual shopping budget on this day. Online retailers launch aggressive marketing campaigns dictated in part by previous year's trend as well as relevant current factors.

Our goal in this study is to use a Black Friday consumer spending dataset to come up with a model that best predicts how much someone is likely to spend on that day, provided a general consumer profile. To this effect, we first carry out a detailed study of the dataset (Section II) itself to see what features are available and their general trends. In Section III, we briefly describe the tools used. Section IV details the pre-processing that had to be done to make the dataset digestible for our models. Section V discusses the baseline models that we compare our models against. Section VI and VII discuss the different factors that determined our choice of models and the results obtained from these models respectively. Directions for future work are touched upon in Section VIII while related work is summarized in Section IX. For a look at the code producing the following models and data exploration, see our GitHub repository in our references [7].

II. Data Exploration

A. Understanding the data set

Our data set originally comes from a competition hosted by Analytics Vidhya but is published on Kaggle [1]. It consists of almost 550,000 consumer transactions from Black Friday, and this will be split .66/.33 into training and test data for our models. Each row represents a single transaction by a specific customer identified by a user_id, and it is important to note that some users have multiple transactions. The raw data is structured as follows:

TABLE I
Raw consumer data features and types

user_id	product_id	gender	age	occupation
integer (nominal)	integer (nominal)	M, F	range	category

city	years in current city	marital status	product category 1
A, B, C	1, 2, 3, 4+	0,1	range

product category 2	product category 3	Purchase
range	range	cents

Table I shows the types of values for each attribute. Most of the attributes are self-descriptive, and the product category labels relate to the types of products purchased for each transaction. Some of the product category 2 and category 3 labels are missing, so we account for that in the pre-processing step. Additionally, we had to binarize some of the categorical variables. During the initial stages of data processing, we explored the data to gain a better understanding of the features and trends. The mean purchase cost is $93.33 overall, while the mode is $68.55. Roughly a quarter of the consumers identify as female, and men spent an average of $7.00 more per purchase than women. The 51-55 age group has the highest average spending, and consumers in city category C spent roughly 10% more than consumers in other city categories. The data set can be found for download here [1].

B. Decision tree analysis

We used decision trees to explore cross dependencies between different features in the data set. For example, let us assume that feature X varies proportionally to feature Y if feature Z is greater than 0 and inversely otherwise, then we would like our models to be able to learn this behavior by including the X*Y conjugate feature. Decision trees are really useful in analyzing the variations in feature values based on any number of dependency relationships.

We trained a decision tree regressor on our model and visualized the result to observe any cross dependencies. A decision tree with the default minimum split value proved difficult to analyze as the dot graph visualization tool would crash after constructing an invalid image file. In order to work around this issue, we had to reduce the split size to 2500 which ensured that any branch with fewer than 2500 samples was not explored further. This yielded a valid tree which was possible to analyze but is still quite dense.

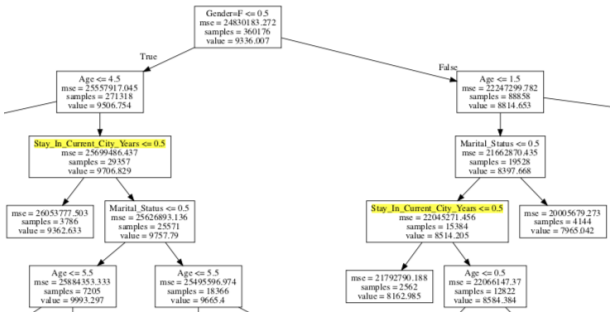

Fig. 1. Visualized decision tree.

Fig. 1 is a snippet of the root of such a tree. The maximum information gain among all user-centric features was from gender. The left-hand side of the tree corresponds to samples with male consumers and female for those on the right side. The *Stay_in_Current_City_Years <= 0.5* corresponds to less than a year spent in the city. We can observe from the figure that regardless of the gender and the age group, more years in the current city implies more money spent on purchases. Similarly, we did not observe any significant fluctuations in the overall trends of how other features varied but there were a few minor ones. Both genders were observed to spend more having lived more years in the same city, however, the trend got slightly reversed for single men. We tried to incorporate this cross-dependency to observe if it made a significant difference in the accuracy of our results but the variations were negligible if at all [8].

III. Tools Used

Pandas [2]: It is an open-source python library providing access to high-performance data structures and algorithms for numerical analysis. We used Pandas DataFrames to store and manipulate our data.

Sklearn [3]: It is one of the many modules of the Scikit library which implements various machine learning algorithms. We use for most of the models used in our code, including decision trees, regressors, etc.

Dot [4]: It is the tool that we used to visualize the decision tree we obtained from the DecisionTreeRegressor in the Sklearn library.

IV. Pre-processing

In order to properly analyze our data we had to represent every entry as a unique feature array. To this end, we explored numerous interpretations and representations of the data. Specifically, as the data wasn't explicit in how the product categories related to one another, this resulted in the brunt of our exploration in representation. First a naive interpretation was utilized wherein the product categories were assumed to be independent of one another, and null product categories were initialized to 0. Note, that this did not result in any conflict or unintentional overlap as the product categories were naturally indexed at 1. From here the product categories along with gender, city, and occupation, were all one-hotted, with user and product id were stripped. This left age and years in city, which were re-scaled such that the magnitude of the feature was proportional with the magnitude of the record. A bias feature initialized to 1 was also included in the vector. This interpretation unfortunately resulted in exceptionally low accuracy across all of our models with an averaged accuracy of .109. While experimenting we attempted to one-hot age and years in city as well, but this did not have a substantial increase or decrease in accuracy. However, when the product categories were concatenated together, and then one-hotted rather than being treated independently, accuracy rose dramatically to an accuracy of .653. This did incur substantially increased overhead as the number of features dramatically rose to take into account every existing combination of product categories. With this insight a more nuanced attempt at interpreting the product categories was attempted, where one feature was product category 1, one feature was the concatenation of 1 and 2, and a third feature was the original concatenation of all 3 product categories. This again multiplied the run time of all tests as the total number of features again rose dramatically, but was not met with a similar increase in accuracy, and in fact resulted in no noticeable accuracy increase. As a result, this interpretation was scrapped and reverted back to solely using the concatenation of all 3 product categories.

The high importance and correlation with product categories observed through the comparison between proper interpretation and naive interpretation was further backed by ablation tests. Ablation tests found that removing any feature other than the aggregate product category had exceptionally little impact on accuracy, with maximum variance between any two tests being .0015. However, ablating the aggregate product category resulted in an accuracy of .011 implying that the other features are almost conditionally independent of the expected purchase price given the other features.

V. Baseline Results

We interpreted the mean as the result of all predictions and used it as our baseline model. Using a central tendency measure provides a basic estimation of a consumers Black Friday purchase amount and is a common baseline measure for linear regression. The mean purchase total of the training set was $93.33, while the mode was $68.55. Over the test data, the R-squared score, which statistically measures how close the data are to the fitted regression line, is 0.0. Since the score is zero, this indicates that the model explains none of the variability for the data. Intuitively, this makes sense; the feature input for each transaction has no bearing on the resulting purchase total because all predicated purchase totals are equal. Yet this baseline provides us with a good estimation for an average persons purchase total on Black Friday. "Fig. 2" below shows the predicted results graphed against actual purchase total.

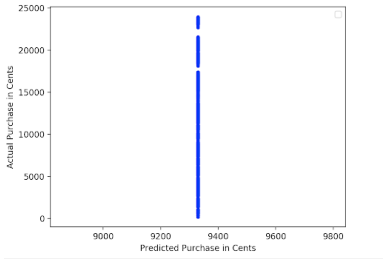

Fig. 2. Baseline predicted purchase results fitted against actual purchase total.

VI. Model Choices

A. Least Ordinary Squares

As an introductory model, we went with a model implementing least ordinary squares, which also provides another baseline fit of the data against a linear model. Least ordinary squares uses coefficients to minimize the sum of squares between responses in the data set [8].

B. Ridge Regression

Ridge regression is a technique that addresses some of the problems created by least ordinary squares. It reduces the standard error by adding a a degree of bias to the regression estimates, thereby imposing penalties on the size of coefficients. Alpha is the input parameter that controls the amount of shrinkage. We then tried this model to see if we could improve upon the more basic linear model using LOS. There is a trade off though since a higher alpha leads to greater bias but lower variance [8].

C. Lasso Regression

Lasso regression uses both regularization and variable selection methods to try to improve the overall fit of the model, so it becomes useful in estimating sparse coefficients. Since some of the product category data was sparse and we removed several features in the pre-processing stage, we thought this technique could provide unique improvements to the model. This would take the model one step farther by accounting for this sparsity [8].

D. Kernel Ridge

We then tried using kernel ridge regression which combines ridge regression with the kernel trick. To fit against a non-linear curve, we used RBF kernels with varying gammas. Kernel ridge uses squared error loss, and can take longer on larger data sets than SVRs, which is a regression application of SVMs [8].

E. Random Forest Regression

We then wanted to try an ensemble model, so we used a random forest regressor which uses multiple decision trees and bootstrap aggregation to try and improve the predictive accuracy. Since it uses bagging to create sub-sample models, this also helps control over-fitting. Random forests also look at other non-linear models and compares scores across a wide range of models for the overall best fit [8].

VII. Model Analysis

With regards to our linear models, lasso, ridge, and least ordinary squares, it probably shouldn't come as much of a surprise that they have exceptionally similar accuracy both on training and test, as shown below in "Table II".

TABLE II
Model Results

Model	***Score(training)***	***Score(test)***
Decision tree, 2500 min-split	.6538	.6515
Least ordinary square	.6526	.6526
Ridge regression, alpha = .5	.6526	.6526
Lasso regression, alpha = .1	.6524	.6524
Random forest regression	.8301	.5924

R-squared scores over training and test data

This does indicate however that given the current representation of the data, and more specifically how the product categories were interpreted, it's unlikely that any linear model will be able to demonstrate dramatic improvements. However, of potentially more interest is that although all three models converged to a similar estimation of values, to the extent that they even had similar outliers and failed to adequately evaluate many of the same entries in our data, they still had relatively dramatic differences in weights given to various features. For instance, the least ordinary squares model noted a slight positive correlation with marital status that was not noted by either of the other models. In reality, based on the data, there is indeed an exceptionally small increase in the amount that married individuals spend over unmarried individuals. This suggests that perhaps the least ordinary squares model was marginally more sensitive to small patterns, and this in turn could explain the minor increase in score of it over and lasso. Although they had their differences, all models ended up having their largest weights attached to the various product categories. This corroborates the ablation testing done earlier that indicated that the categories played the largest

roll in determining the purchase price. Even though all three models did have their highest weights associated with the products categories, there were significant variations in how the individual product categories were evaluated here too. Least ordinary squares had many of the categories at the same weight with only a few notable exceptions. Ridge alternatively had the largest variance between product categories, and finally lasso was unique in that it found several category compositions that it believed to have no weight at all. As all three models still converged to give very similar results, its likely that there are some deeper roots in how the product categories interact with the other attributes that can't be properly captured in linear a model, but allow for multiple different weight assignments to still arrive at similar conclusions. "Fig. 3" below shows the predicted total versus actual purchases total for these three linear regressors.

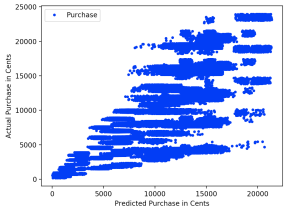

Fig. 3. LOS, Ridge, and Lasso linear regressors.

Since the random forest regressor is an ensemble model that utilizes bagging and fits multiple models, including non-linear models, it provided a more wide spread attempt at maximizing the R-sqaured score across the training and testing sets [8]. We experimented with both the max depth and number of estimators and found that having too high of a max depth led to over-fitting of the training data, resulting in a poor score for the test set. Increasing the number of estimators, or trees in the forest, improved the test set score slightly, but resulted in longer run times since the forest was expanded. After tuning those parameters, we fit the model against a sub-sample of the original data set for complexity reasons, and the predicted purchase total mapped against the actual purchase for the testing set is show below in "Fig. 4"

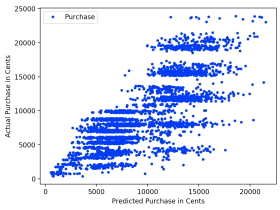

Fig. 4. Random forest predicted purchase.

On the testing set, the model scored roughly .6, meaning that there is a noted correlation in feature values to predicted purchase, and building a random forest regressor on a larger subset of the data results in a higher score on the testing set. As you can see in "Fig. 4", most data points are centralized around a linear line of fit, and the distance of the point from this line represents how inaccurately the purchase total was predicted. The random forest far exceeded the other models on the training set score, however, with a score of almost .85 continuously.

Unfortunately, running both the RBF SVR model and kernel ridge regressor with a RBF fit ran for either hours or utilized all the memory of the server before completion. The hope was to use these non-linear models to compare the fit with the linear models described above. Since the kernel ridge regressor does not operate on sparse matrices, it attempts to convert them to standard arrays by allocating an all-zero array and then filling in the appropriate spots with the values sparsely stored, likely resulting in the enormous memory usage [9]. SVR's are typically faster on larger data sets, yet than kernel ridge [8], but with the available servers, running the model timed out before successfully fitting the data set.

VIII. What's Next

Given more time and space, we would have liked to expand the scope of this project to gain further insights. We tried to experiment with non-linear models but ran into various issues. With our implementation of RBF kernel ridge model, our program crashed with complaints of running out of memory. This is a well-documented issue [9] but we were unable to come up with a working solution by the time of writing of this report. Given more time we would have liked to make use of a large-scale machine learning framework such as Tensorflow [10] to build more complex non-linear models to observe the effect on model accuracy, if any.

Another insight moving forward could be attempting to classify the gender of a consumer based on other demographic and purchase features. This would open the doors to experiment with classifiers like k-th nearest neighbor, neural nets, and naive Bayes.

IX. Related Work

Stewart et al. [5] predict consumer spending as obtained using a Gallop poll based on the volume of Twitter keywords used by various users. Their approach offers better accuracy over models exclusively based on current user spending trends. Ming et al. [6] carry out an extensive study to investigate spending patterns and online browsing behavior and predict overall spending patterns on one of the big Chinese holidays. They make use of collaborative filtering and cross-validation approaches to further explore critical shopping behaviors.

References

[1] Mehdi Dagdoug, "Black Friday: A study of sales trough consumer behaviours" https://www.kaggle.com/mehdidag/black-friday
[2] McKinney, Wes. "Pandas: a foundational Python library for data analysis and statistics." Python for High Performance and Scientific Computing 14 (2011).
[3] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830
[4] Gansner, Emden R., et al. "A technique for drawing directed graphs." IEEE Transactions on Software Engineering 19.3 (1993): 214-230.

[5] Stewart, Justin, et al. "Twitter keyword volume, current spending, and weekday spending norms predict consumer spending." 2012 IEEE 12th International Conference on Data Mining Workshops. IEEE, 2012.
[6] Zeng, Ming, et al. "User behaviour modeling, recommendations, and purchase prediction during shopping festivals." Electronic Markets (2018): 1-12.
[7] Nora Myer, Aidan Globus, Haseeb Javed, "Black Friday linear regression analysis" https://github.com/noramyer/black-friday-lin-regression
[8] Sci-kit Learn Organization, "User guide". (2018). https://scikit-learn.org/stable/user_guide.html
[9] StackOverflow, "Python MemoryError when doing fitting with Scikit-learn". (2018).https://stackoverflow.com/questions/16332083/python-memoryerror-when-doing-fitting-with-scikit-learn
[10] Abadi, Martn, et al. "Tensorflow: A system for large-scale machine learning." 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). 2016.