PREDICTIONS ON

BLACK FRIDAY DATASET

EAS 503, Fall Semester
Final Project Report

Project Members:
Akhila Devabhaktuni
Pratik Sanghvi
Yao ji

Reported to:
Varun Chandola,
Professor Computer Science and Engineering
SUNY at Buffalo

Table Of Contents

Introduction:

In our project, we wanted to take the opportunity to explore sql database integration in python, relational algebra for data manipulation, using scikit-learn package to fit different models to data and developing an interactive dashboard.

At this time and age, target marketing had become a very integral part of every company's profit strategy. So, we've picked Black Friday Sales dataset from kaggle. Our main reason for choosing this dataset is,it's more close to the data collected in real time and serves our goal of exploring various algorithms in regression and classification analysis for various categories of target customers.

Our development strategy contains in part of selecting the relevant dataset, by searching for a reliable source, getting the data into our local database server, followed by a careful study of data, cleaning by removing or filling the missing components followed by Exploratory data analysis to find correlation of variables in data.

Our end goal is to observe data trends by applying various models(Multiple linear regression, Ridge etc.) to the data available depending on the variable examined. And to do so using an interactive dashboard with the help of real time data.

We've employed *mysql* package of python for data extraction, cleanup and subsetting, then based on the chosen algorithm and end variable, using *sklearn* package of python, various observations are made and presented on the dashboard.

Data Description:

Our dataset consists of 550000 observations and 12 variables. These variables are of either categorical or numerical. The variables include:

Age**:** Categorical variable having 7 categories of age_groups.

Gender: A categorical variable with values of M and F.

Occupation: Numeric variable having 20 occupation types. Though numeric we have treated it as a categorical variable.

City_Category: A categorical variable with values of A, B and C. We treat these as Tier_1, Tier_2, Tier_3 cities respectively.

Stay_in_Current_city_Years: A categorical variable with values of 1, 2, 3 and 4+ years.

Marital_Status: A numeric variable with values of 0 and 1, we treat it as a categorical variable with 0 indicating Single and 1 indicating Married

Product_Category_1: A numeric variable with 18 values, we treat it as a categorical variable having 18 categories of products

Product_Category_2, Product_Category_3: We drop these variables from our analysis

Purchase: A numeric variable having having information about the purchase amount of each transaction.

Dashboard Description:

Our dashboard consists of 2 sections. The first one involves looking at the historical data and doing a exploratory data analysis using different visualization. And the second one has all the models for predictions.

Exploratory data analysis section involves the graphs that shows the interaction between the predictors to get a sense of the data. It has histograms and boxplots showing the interactions with our target variable Gender. This tab gives us a general idea about the predictors, by looking at them we can make inferences on the inclusion or exclusion of variables as predictors in our model.

Packages Used:

We've employed various packages from Python discussed in class for our data manipulation and model fitting purposes.

Numpy, Pandas, ipywidgets, IPython, pymysql, matplotlib, warnings, seaborn, scikitplot, scipy

Sklearn has many learning algorithms, for regression, classification, clustering and dimensionality reduction.

Data Import and Data Cleaning:

Our dataset was in the form of a csv file. We created our database in mysql. We created a number of relations that would be joined with
our main dataset.

We had a relation called age_groups that had a shorter category of age groups and a broader category (youth, working class and
Senior Citizen).

One relation named Product_Categories had the description of 18 product categories and Occupation had 20 occupation descriptions.
The joining of these tables have been done with pymysql package in python.

Regression Models:

Studying purchase strategies of customers will surely help in customized coupon generation and better service. To predict purchase value of a customer, various aspects are observed and fit into the model chosen. We've divided our dataset into test and training to fit the model and optimize it. We've generated a validation dataset for error calculation.

a)**Multiple Linear Regression**:

The method works on simple estimators as well as on nested objects (such as pipelines). We've performed this using sklearn library and training the model by 'LinearRegression' from sklearn library.
Since some fr ur columns are categorical we need to transform them first as string.

Now, we have the data frame that contains the independent variables and the data frame with the dependent variable (marked as "purchase"). Let's fit a

regression model using SKLearn. First we'll define our X and y , divide them into test and training and fit a linear model.

After we have a model, predict test and training mean squared errors for accuracy calculations.

```
Coefficients:
 [ 529.04632447  114.75842763    6.43493161  369.21665288   12.25116013
  -49.55015643 -412.58157725]
Mean squared error: 51.68
Variance score: 0.10
```

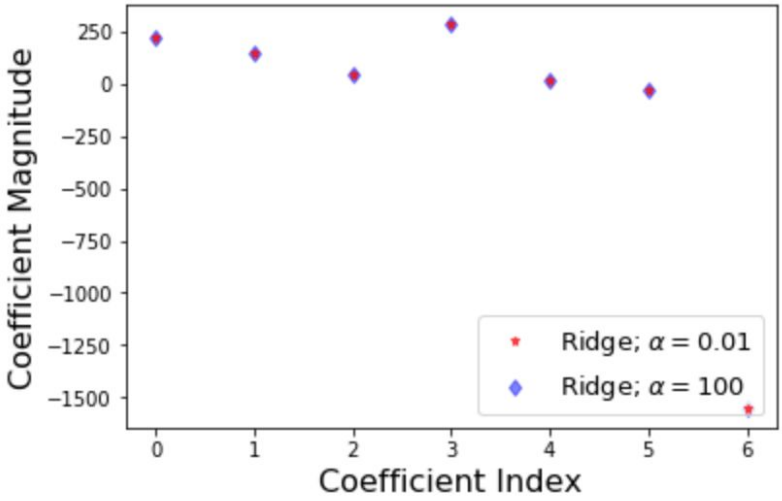
We can also observe the coefficients of various variables to know which variables influence the end results more.

b)**Bayesian Ridge Regression**:

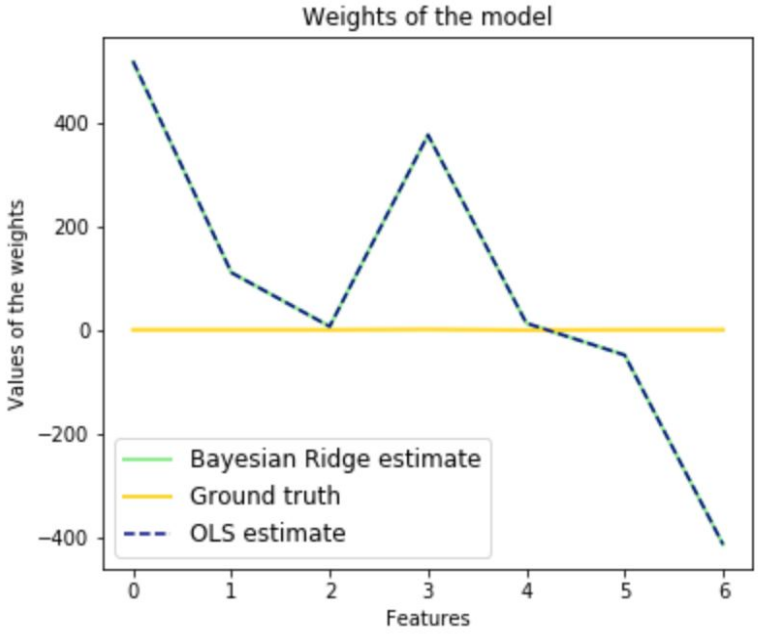
Used to estimate a probabilistic model of the regression problem. Compared to the OLS (ordinary least squares) estimator, the coefficient weights are slightly shifted toward zeros, which stabilises them.The histogram of the estimated weights is Gaussian.The estimation of the model is done by iteratively maximizing the marginal log-likelihood of the observations.

Dividing the data into test and training and fitting model is same as before, except the model itself changes.

Changing the value of alpha didn't make a significant difference to the model.

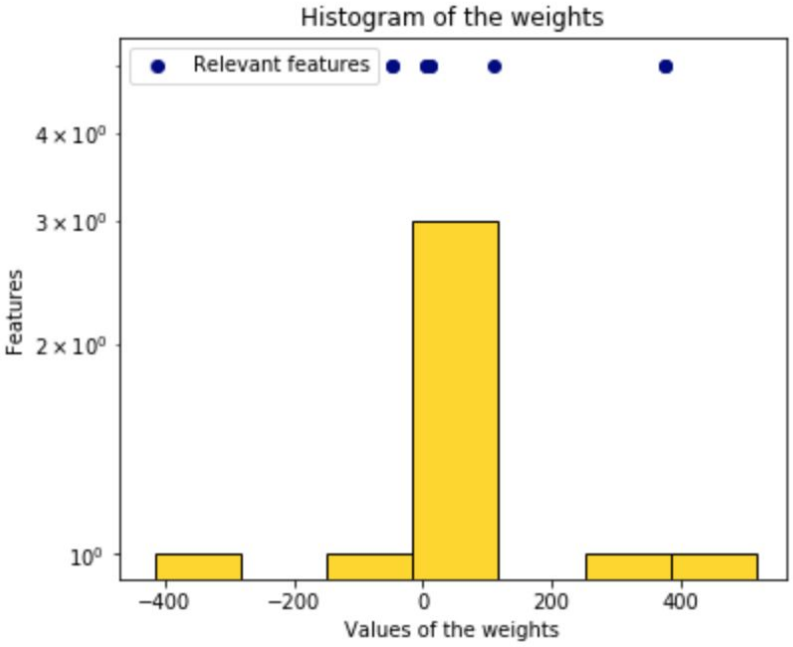


By setting the ground truth to zero, performance of Bridge regression compared to Ordinary least squared can be observed from the following graphical representation

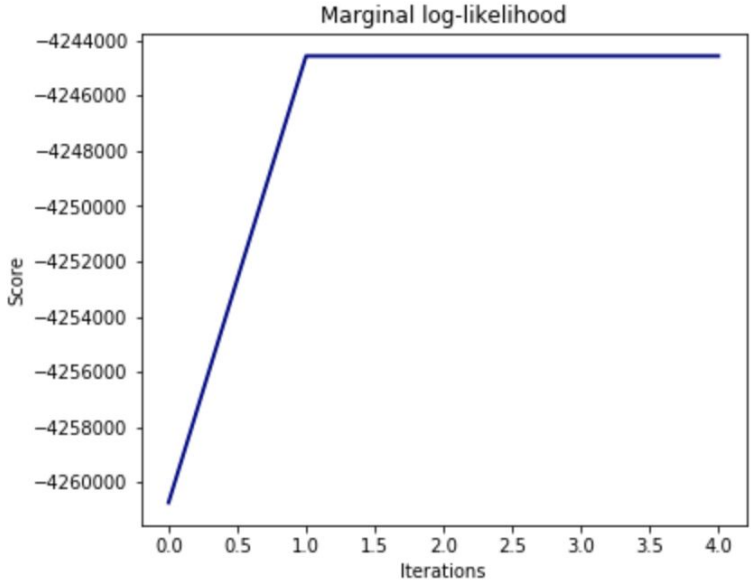


.

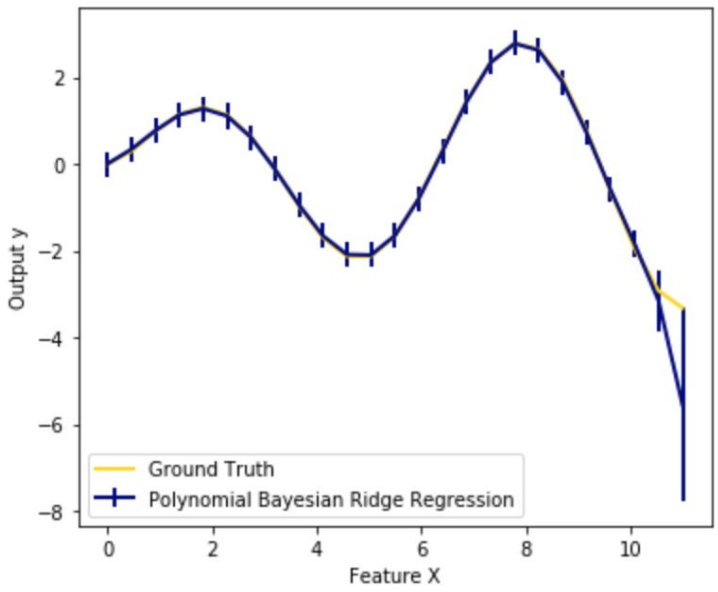
And to understand the relevancy of weights, we plot a histogram and observe the variable performance in context of the rest.



Marginal Likelihood graph

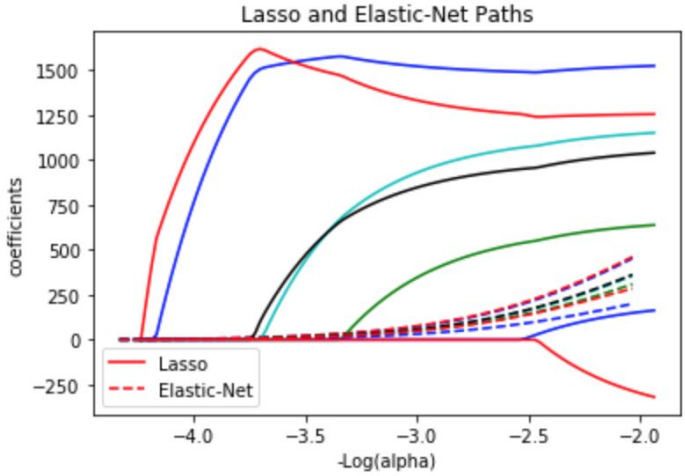


On observing the performance of Polynomial Bayesian Ridge Regression when compared to the ground truth:

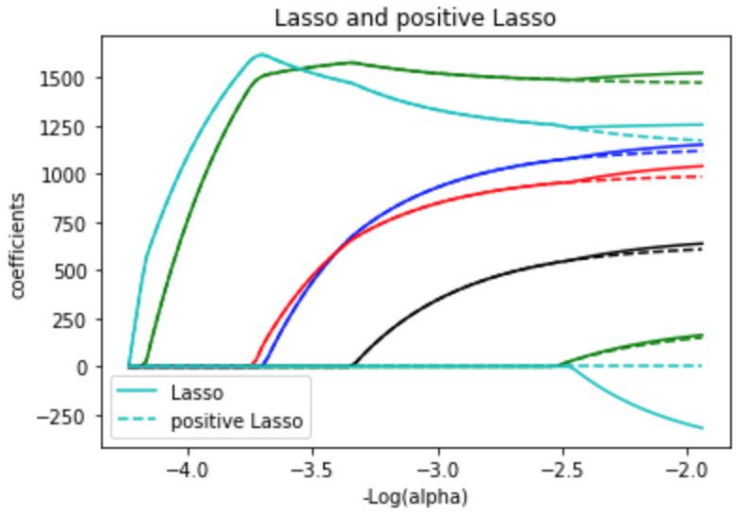


c) Lasso and Elastic net paths:

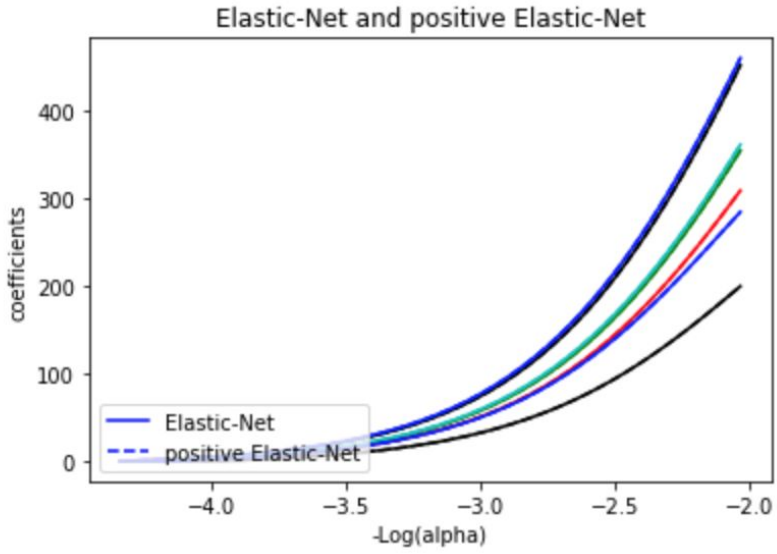
Just to have an understanding of when a variable loses its importance over time.



Lasso and positive lasso graphs:



Elastic net and positive elastic net:



Classification Models:

In classification models, we chosen “Gender” as our categorical dependent variable that has two classes “F” and “M”. In order to obtain and compare the accuracy of the prediction, we selected four classifiers: Logistic regression, Decision Tree, K Neighbors (KNN) and Linear Discriminant Analysis (LDA) as our classification models to fit the training data and predict the probability of each class of “Gender” in test set. We also used the precision-recall curve method to predict the precision of each class of “Gender”.

Precision-recall Curve Parameters:

Precision is defined as the number of true positives (tp) over the number of true positives plus the number of false positives (fp).

Recall is defined as the number of true positives (tp) over the number of true positives plus the number of false negatives (fn).

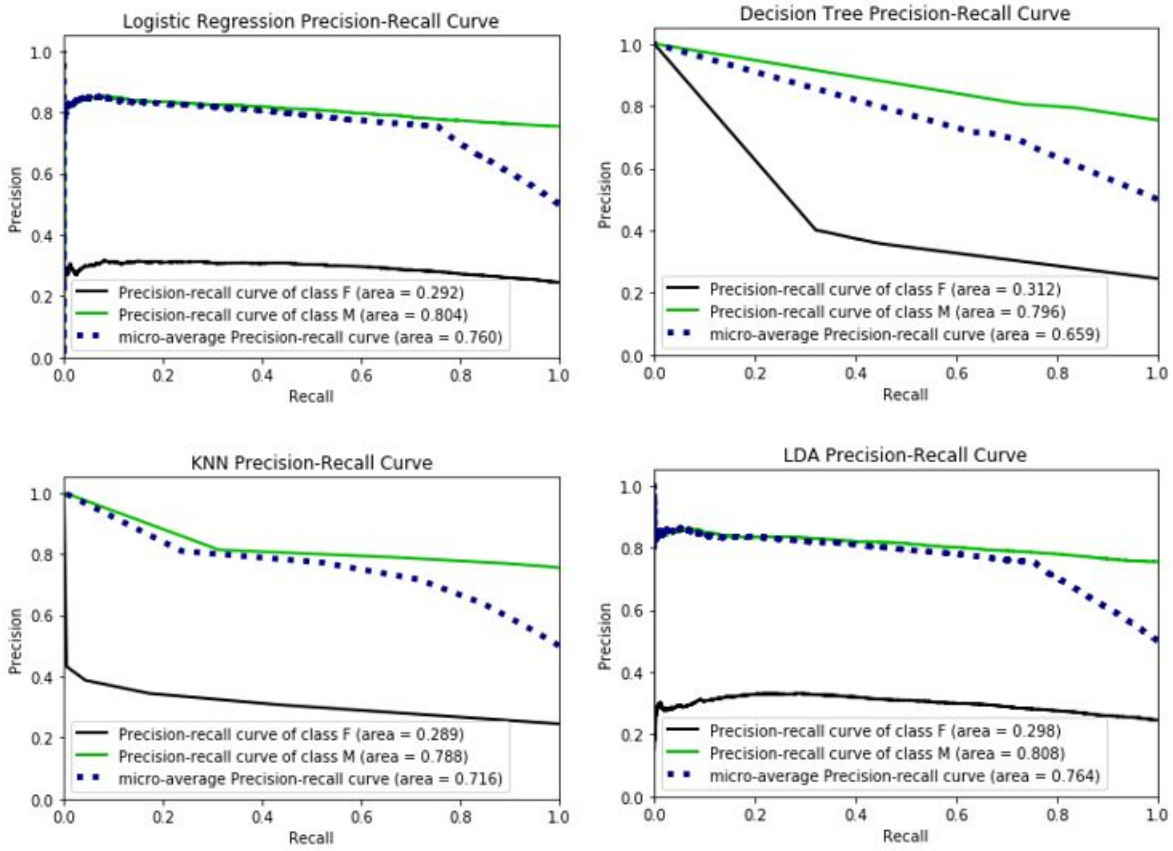
$$Precision = \frac{tp}{tp+fp} \text{ and}$$

$$Recall = \frac{tp}{tp+fn}$$

where **tp = True Positives, fp = False Positives and fn = False Negatives**

Outputs:

Classification Models	Accuracy	Precision-recall curve avg area	First 10 test results
Logistic regression	0.76	0.76	['M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M']
Decision Tree	0.68	0.66	['F' 'F' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M']
KNN	0.72	0.72	['M' 'M' 'F' 'M' 'M' 'M' 'M' 'M' 'M' 'M']
LDA	0.76	0.76	['M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M']



Based on the results above, logistic regression model and LDA have more higher accuracy than KNN and Decision tree models. In addition, the precision-recall curve of various models shows the AUC (area under the curve) of each class of dependent variable “Gender”, we found that the AUC of Male appeared in test set is much higher than that of Female.

Observations & Insights:

Using the visualizations, we observed that the data is biased towards Male, around 75% of the data is for Male and 25% is Female, so if we make a prediction for gender using this data, the prediction will be biased towards Male. This is one of the important insights we could get. For future prospects we would have to use techniques like oversampling and undersampling and then use the data for making predictions.

We've used Lasso, positive Lasso, Elastic Net and positive Elastic Net to understand the behaviour of variables and their relative importance over varying iterations.

Some typical observations include place of residence of an individual seem to have no considerable effect on purchase value. Most influential variables are an individual's occupation status, age and gender.