# Sales Forecasting Using Machine Learning Techniques

## A Case Study of Regression Algorithms with Categorical Variables

Richard Liu[1]
[1]Allen High School, 300 Rivercrest Blvd, Allen, TX 75002, USA

***Abstract*—In this paper, we use a public dataset from a retail store to forecast purchases on Black Friday. A variety of regression algorithms, including linear regression, decision trees, random forest, and gradient boosting have been applied to make predictions. In addition, we discuss how to handle categorical predictor variables in the dataset. The comparison on the accuracy of algorithms is performed and a detailed analysis on the results is provided. Furthermore, we investigate the impact of categorical predictor variables in the dataset on prediction performance.**

***Keywords—Supervised Machine Learning; Regression; Categorical Variables***

## I. INTRODUCTION

The retail business has become more and more competitive. Applying cutting-edge machine learning techniques can not only help businesses analyze the substantial and complicated data to learn sale patterns, but also predict future sales to optimize supply chains and further stay ahead of their competition. Some successful use cases on utilizing machine learning to boost retail business have been reported [1, 2].

Although machine learning is revolutionizing the retail business, it also faces challenges for multiple reasons: (1) missing data and outliers, (2) too many features/attributes, and (3) features that do not describe the products.

In this work, using public data from the "Black Friday" Kaggle dataset [3], we apply various algorithms including linear regression, decision trees, random forest, and gradient boosting to perform predictions. We first fill in the missing data. Then, different approaches were conducted to handle categorical variables in the dataset before feeding the variables into regression algorithms.

The rest of this paper is organized as follows: in Section II, we provide an overview on related work. In Section III, we describe the dataset used for this work. In Section IV, we evaluate various regression algorithms, compare the prediction performance, and provide analyses on accuracy performance. Section V concludes the paper and points out future work.

## II. RELATED WORK

A considerable amount of literature has focused on predicting sales in the retail business. S. Sharma and V. Sharma presented a comparative analysis on machine learning techniques in sale forecasting and proposed using both moving averages and artificial neural networks with backpropagating to achieve higher accuracy [4]. The authors in [5] provided a survey on both statistical and machine learning (neural networks) techniques used to forecast sales in fashion markets. In [6], the authors applied an extreme learning machine (ELM) to investigate the relationship between sales and the attribute factors in fashion retailing. Besides the fashion domain, Y. Kaneko and K. Yada proposed a deep learning approach for general retail store sales [7]. For the housing market, B. Park and J. Bae develop a housing price prediction model based on the RIPPER algorithm [8].

## III. DATASET AND METHODOLOGY

### A. DataSet and Problem Statement

We use a public dataset named "Black Friday" from Kaggle (full details at Kaggle [3]), which includes 550000 transactions in a retail store. It contains various customers' information (e.g. age, gender, occupation, etc.) and the number of purchases for each customer on Black Friday. A sample of the dataset is shown in Figure I. As seen, there are 12 predictor variables in the dataset, as shown in Table I.

TABLE I PREDICTOR VARIABLES IN DATASET

| Variable | Definition |
|---|---|
| User_ID | User ID |
| Product_ID | Product ID |
| Gender | Gender of user |
| Age | User age in bins |
| Occupation | User occupation |
| City_Category | Category of the city |
| Stay_in_Current_City_Years | Number of years the user has stayed in current city |
| Marital_Status | User marital status |
| Product_Category_1 | Category (may belong to more than one category) |
| Product_Category_2 | Same as above |
| Product_Category_3 | Same as above |

TABLE II AGE BIN MAPPING

| Age Range | 0 - 17 | 18 - 25 | 26 - 35 | 36 - 45 | 46 - 50 | 51 - 55 | 55+ |
|---|---|---|---|---|---|---|---|
| Mapped Age | 15 | 21 | 30 | 40 | 48 | 53 | 60 |

The target variable in the dataset is Purchase.

Given the features in the dataset, our objective is to investigate a host of regression algorithms to predict the purchase value after necessary preprocessing.

### *B. Data Preprocessing*

Preprocessing is necessary because the dataset contains several categorical variables, including Gender, City_Category, Marital_Status, and Age. In addition, there are some missing values in Product_Category_X.

K. Potdar, T. S. Pardawala and C. D. Pai performed a thorough comparison on various encoding schemes for categorical variables [9]. Since the categorical variables in our dataset are either binary or with a limited number of values, we use the most common technique, one hot encoding, to preprocess the data. Specifically, one hot encoding is used on Gender and Marital_Status due to their binary characteristics. In addition, since City_Category only has three values: A, B and C, one hot encoding is also used here to encode the City_Category variable. For Age, a mapper is applied to convert numeric bins to a specific value as shown in Table II. In data preprocessing, we rename Age to Age_Encoded, which will be used in regression algorithms.

Although Occupation can be treated as numeric variable, its value is just represented as an ID. Therefore, its magnitude does not indicate any ordinal relationship. For this variable, two scenarios are considered: (1) treating it as a numeric variable and directly feeding it into regression algorithms, and (2) using one hot encoding first before feeding it into regression algorithms.

The missing values in Product_Category_X are just filled with zero, assuming the products do not fall into that category.

## IV. IMPLEMENTATION AND EVALUATION

### *A. Exploratory Data Analysis (EDA)*

The correlation between the features and purchases is calculated first and results are shown in Figure II.

As expected, Age (Age_Encoded) and Marital_Status (Marital_Status_1) have the strongest correlation. In addition, although Age and Occupation do not show a strong correlation, the distribution of the Occupation category in each age group still aligns with employment profiles. As shown in Figure III, young age groups (e.g. ages 0-17 and ages 18 - 25) can only take very limited types of jobs due to having smaller skill sets, so there is only one large spike in the occupation distribution. With age increasing, the occupation distribution becomes more diverse. In addition, the spikes in the senior groups are different from those in the junior groups, indicating that most of the seniors take different types of jobs than the juniors take.

Regarding the Purchase target variable, Product_Category_1 shows a negative correlation with Purchase while Product_Category_3 has a strong correlation with Purchase, implying that the products in Category 3 may be on sale. Other than these two variables, no other variables are highly correlated with Purchase.

### *B. Implementation*

To forecast Purchase, we apply different regression algorithms to perform predictions and evaluate their performance. Specifically, linear regression, decision tree, random forest, and gradient boosting are chosen. Python sklearn and pandas libraries are used to import the dataset, process the data, and make predictions. In all cases, the dataset is split into two subsets: 80% of data for train samples, and 20% for test samples.

As mentioned in Section III-B, for each algorithm, we further investigate two scenarios: (1) Occupation variable is directly fed into algorithms, and (2) Occupation variable is first encoded using one hot encoding before applying algorithms.

### *C. Evaluation*

For each algorithm, Root Mean Squared Error (RMSE) and R-squared value of the prediction are calculated and serve as performance metrics.

#### *C.1 Occupation - numeric*

Table III below shows the prediction performance for each algorithm.

TABLE III PREDICTIONS ON NUMERIC OCCUPATION

| | Linear Regression | Decision Tree | Random Forest | Gradient Boosting |
|---|---|---|---|---|
| RMSE | 4627.69 | 3275.32 | 3076.10 | 3076.10 |
| $R^2$ | 0.13 | 0.57 | 0.62 | 0.64 |

#### *C.2 Occupation – one hot encoding encoded*

Applying one hot encoding on Occupation outputs 20 variables. To reduce the dimensions of the feature space, PCA (8 components) is used after the dataset is normalized.

Table IV below shows the prediction performance for each algorithm.

TABLE IV PREDICTIONS ON ONE HOT ENCODING OCCUPATION

| | Linear Regression | Decision Tree | Random Forest | Gradient Boosting |
|---|---|---|---|---|
| RMSE | 4631.95 | 3215.41 | 3074.39 | 3076.39 |
| $R^2$ | 0.13 | 0.58 | 0.62 | 0.61 |

### D. Result Analysis

The results from both cases show that linear regression has the worst prediction accuracy. This indicates that the linear relationship between the independent and dependent variables is weak. In addition, linear regression assumes the data are independent. However, the EDA in Section IV-A shows that some data (Age and Marital Status, Age and Occupation) have some correlation, which contributes to the degraded prediction accuracy.

The three other algorithms (decision tree, random forest, and gradient boosting) can capture the non-linear relationship in the dataset; therefore, they achieve better prediction accuracy.

For our evaluation, all regressors use default values for each parameter in the algorithms. In the decision tree case, the maximum depth of the tree is set to None, which means "nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split" [10]. Due to this, the decision tree may overfit, which explains why its accuracy is worse than that of random forest and gradient boosting.

Random forest and gradient boosting achieve about the same accuracy in both cases (one hot encoding and numeric Occupation). The main difference between the two algorithms lies in the way the trees are built: random forest trains each tree independently using a random sample of the data, while gradient boosting builds trees one at a time, and each new tree helps correct errors made by the previously trained tree. Theoretically, after fine tuning algorithm parameters, gradient boosting should perform better than random forest. However, we use default values in both algorithms, which leads to highly similar accuracy.

One interesting observation is that using one hot encoding on the Occupation predictor variable does not improve the prediction accuracy as expected. There are several reasons why. Firstly, to reduce dimensions of feature space, we set 8 components in PCA. This may be not an optimal number, which may compromise prediction accuracy. Secondly, the Purchase distribution among Occupation shown in Figure IV indicates that the Purchase is more or less the same for all occupations. Thirdly, Occupation does not seem to be a statistically important predictor variable in regression.

## V. CONCLUSION

In this paper, four different regression algorithms have been used to forecast the purchase for a retail store. We preprocess the data by encoding the categorical predictor variables, performing exploratory data analysis, and then applying algorithms on the dataset.

The evaluation shows that random forest and gradient boosting achieve the same level of prediction accuracy, followed by decision tree. Due to the existence of a non-linear relationship between the predictor variables and the target variable, linear regression bears the worst prediction accuracy.

As further work, we intend to optimize parameters in the algorithms for one hot encoding on the Occupation case to see if better performance can be achieved. In addition, we plan to use other algorithms, like support vector regression, deep neural network regression, etc. on the dataset to perform more comparisons and analyses.

## REFERENCES


[1] B. Randolph, "How 6 Brands are Using Machine Learning to Grow Their Business," https://www.shopify.com/retail/how-6-brands-are-using-machine-learning-to-grow-their-business, Dec 7, 2017.

[2] I. Bobriakov, "Top 10 Data Science Use Cases in Retail", https://medium.com/activewizards-machine-learning-company/top-10-data-science-use-cases-in-retail-6483accc6042, Jul 22, 2018.

[3] M. Dagdoug, "Black Friday: A study of sales through consumer behavious", https://www.kaggle.com/mehdidag/black-friday.

[4] S. Sharma and V. Sharma, "Comparative Analysis of Machine Learning Techniques in Sale Forecasting ", International Journal of Computer Applications (0975 – 8887), Volume 53– No.6, September 2012, pages 51-54.

[5] S. Beheshti-Kashi, H. R. Karimi, K. Thoben, M. Lutjen and M. Teucke, "A survey on retail sales forecasting and prediction in fashion markets", Systems Science & Control Engineering, 3:1, 2015, pages 154-161.

[6] Z. Sun, T. Choi, K. Au and Y. Yu, "Sales forecasting using extreme learning machine with applications in fashion retailing", Decision Support Systems, Volume 46, Issue 1, December 2008, pages 411-419.

[7] Y. Kaneko, K. Yada, "A Deep Learning Approach for the Prediction of Retail Store Sales", IEEE 16th International Conference on Data Mining Workshops, 2016.

[8] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data", Expert Systems with Applications, Volume 42, Issue 6, 15 April 2015, Pages 2928-2934.

[9] K. Potdar, T. S. Pardawala and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifers", International Journal of Computer Application (0975 – 8887), Volumn 175 – No. 4, October 2017, pages 7 – 9.

[10] Scikit learn document, https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html.

| User_ID | Product_ID | Gender | Age | Occupation | City_Category | Stay_In_Current_City_Years | Marital_Status | Product_Category_1 | Product_Category_2 | Product_Category_3 | Purchase |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1000001 | P00069042 | F | 0-17 | 10 | A | 2 | 0 | 3 | | | 8370 |
| 1000001 | P00248942 | F | 0-17 | 10 | A | 2 | 0 | 1 | 6 | 14 | 15200 |
| 1000001 | P00087842 | F | 0-17 | 10 | A | 2 | 0 | 12 | | | 1422 |
| 1000001 | P00085442 | F | 0-17 | 10 | A | 2 | 0 | 12 | 14 | | 1057 |
| 1000002 | P00285442 | M | 55+ | 16 | C | 4+ | 0 | 8 | | | 7969 |
| 1000003 | P00193542 | M | 26-35 | 15 | A | 3 | 0 | 1 | 2 | | 15227 |
| 1000004 | P00184942 | M | 46-50 | 7 | B | 2 | 1 | 1 | 8 | 17 | 19215 |
| 1000004 | P00346142 | M | 46-50 | 7 | B | 2 | 1 | 1 | 15 | | 15854 |

FIGURE I. EXAMPLE OF BLACK FRIDAY DATASET

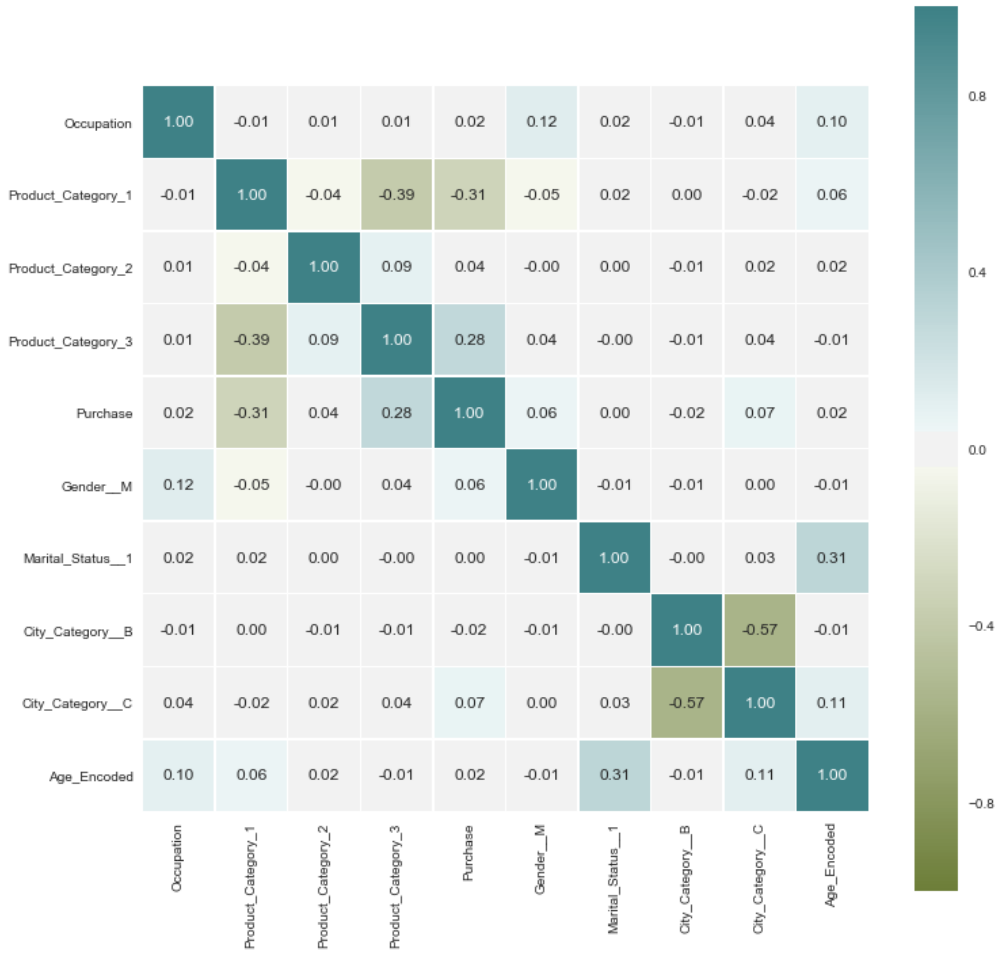


FIGURE II. EDA – CORRELATION

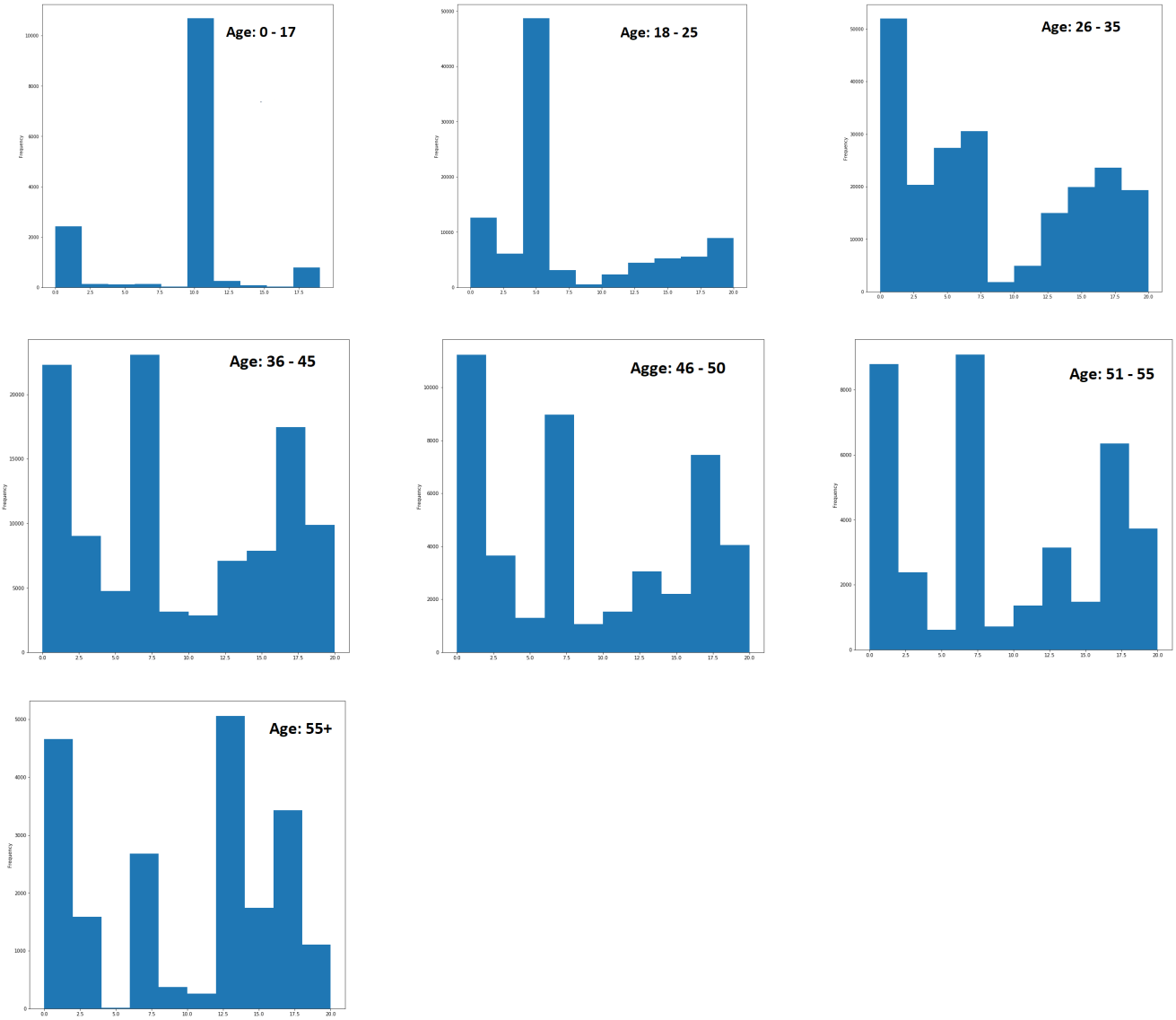


FIGURE III. AGE BIN VS. OCCUPATION (X AXIS: OCCUPATION, Y AXIS: FREQUENCY)

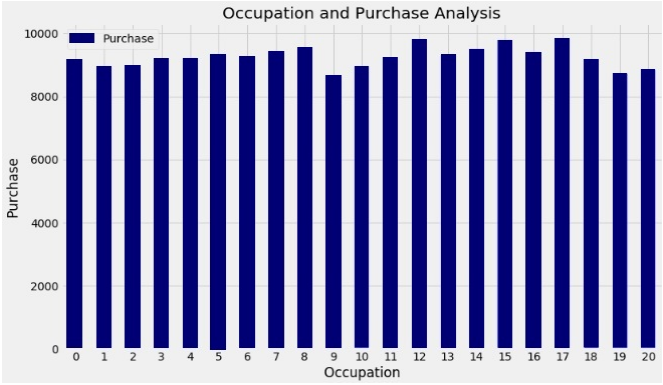


FIGURE IV. OCCUPATION VS PURCHASE