Using Machine Learning to Predict Costumer Ratings of Cereal Products

Fatih Özdemir*
No: 161180753

Ramis Yüksel*
No: 151180072

Eda Nur Subaşı*
No: 151180051

Giray Serter*
No: 161180758

CENG 313
Introduction to Data Science
Project Assigment

December 10, 2019

Abstract

Providing a decent customer satisfaction has always been a enormous trouble for companies, especially for cereal manufacturers. The gathering of the customer reviews has become a huge importance for the market sales.Therefore, predicting potential customer ratings have huge impact to increase sales and decrease negative feedback. Rating of products that do not come onto the market, cannot be labeled via human supervision. Here, we investigated solving this problem via machine learning methods such as SVR, Linear Regression, Decision Trees, Random Forests. Using a clearly labeled dataset from kaggle.com, we achieved a peak test accuracy of 97.4% using different methods and analyzed their differences.

I. Introduction

Before starting to understand the dataset that we choose to examine and research, very well, we should understand carefully what is data, why we need data science, how to use a data set and also how to convert sets of data to an understandable data circuits.

Before learn about data science, first word that should known is data. Data is a set of gathered and translated information form to use for analysis. Second word is big data. Big data, both structured and unstructured, are large collections and traditional software techniques are very large to control and analysis. Third but not least word is machine learning. Machine learning is a part of artificial intelligence that provides and focus on the development and future of computer programs. The mean reason for machine learning , allow the computers learn by themselves without human intervention and adjust actions accordingly.

Generally, the thing that includes data collection that cannot be solved by hand is data science. Data science is a science that helps us to clearly understand and provide meaningful usage of a big data collection. It uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize and interact with data to create understandable data products. In Data Science process, there are steps that should consider

*Computer Engineering Department, Gazi University, Ankara 06570, Turkey http://mf-bm.gazi.edu.tr/

carefully. These steps can count as; Obtaining data, preprocessing of data and editing of data, creating models, testing models, measuring the performance of model.

After getting information about where we can bring this project to life, we search data sets that should have proper for data analyzing process. In order to extensive data set research, we found a dataset which has ingredients in cereal food product with brand names in it. These ingredients will be our data that we will examined.

During examination of dataset give us opportunity to determine the problem in dataset as prediction of costumer cereal rating. Most of the people in the earth who has opportunity to eat proper foods, mostly eating cereals at their breakfast. This thought give us our problem that can solve with this dataset. In recent years, seventy-seven different cereal brands has come up. These cereals has different ingredients percent and has different taste. In this dataset, according to costumers ratings, cereals rated. We plan to see can we predict possible rating for cereals which has ingredient percentage knowledge.

II. Dataset

A. Dataset Description

Our dataset contains 77 different type of cereals. Each type of cereals have 16 features such as:

- name: *cereal's name*
- mfr: *manufacturer of cereal*
- type: *type of cereal*
- calories: *calories per serving*
- protein: *amount of protein as gram in one serving*
- fat: *amount of fat as gram in one serving*
- sodium: *amount of protein as milligram in one serving*
- fiber: *amount of dietary fiber as gram in one serving*
- carbo: *amount of complex carbohydrates as gram in one serving*
- sugars: *amount of sugar as gram in one serving*
- potass: *amount of potassium as milligram in one serving*
- vitamins: *vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended*
- shelf: *display shelf (1, 2, or 3, counting from the floor)*
- weight: *weight in ounces of one serving*
- cups: *number of cups in one serving*
- rating: *a rating of the cereals (Possibly from Consumer Reports)*

Manufacturer of cereal can include different values as American Home Food Products, General Mills, Kellogg, Nabisco, Post, Quaker Oats and Ralston Purina. These values are shown with their first letter in relevant cell. When examining to type as feature, it can include two value: C for cold and H for hot. For Rest of the features are numerical. We found this dataset from kaggle publicly for using in project.

III. preprocessing

A. Missing Data

The preliminary data set we had about 77 manufacturer with 16 features. However the data was slightly ”porous”, as described by one associate. Some of cereal examples have not all the 15 features described, especially at amount of carbohydrates, potassium and sugar columns. This data values marked by "-1".An amount of nutrition can not be negative, so we can assume these are missing values. We needed a full data set to begin training our model.

A method was devised to fill up negative values. First, we tried replace this cells with mean of the column which is relevant, and tested it. Then, we tried replace this cells with zero, and tested it also. When we analyzed the results we saw that scores was better when we replace it with zero. So, any features which have negative values were filled in the working data set with zero. After this filling process,every cereal name with an missing value was handled.

This gave us a working data set with no dimension lost. No need to delete missing rows

B. Encoding

Although most features were described with a manageable and numeric value, some were described with a categorical value that needed to be converted or deleted. For example, the value entered for mfr, which was a text entry from a consistent list of name of manufacturers, for which there were 7 categories. To make the training easier we gave an integer 0 - 6 for Manufacturer of cereal.

Another feature which binary encoded is the type of cereal(hot or cold). Given the numerical values, 1 for

"hot" , 0 for "cold"

C. Feature Selection

There are some correlations between features. We had to analyse this relation for understanding the data and acquiring more consistent predictions. We used Random Forest algorithm for scaling the relation of other features between "rating". According the *Figure 1* the most important, most correlated feature with the target feature (rating) is "sugar" , then "calories" come. It show us they are the key values. The least important feature is "type". Probably, reason of that is being same (95%) most of values are same in that feature. *Figure 1* shows us the importance power with only one other feature, this is not enough alone.

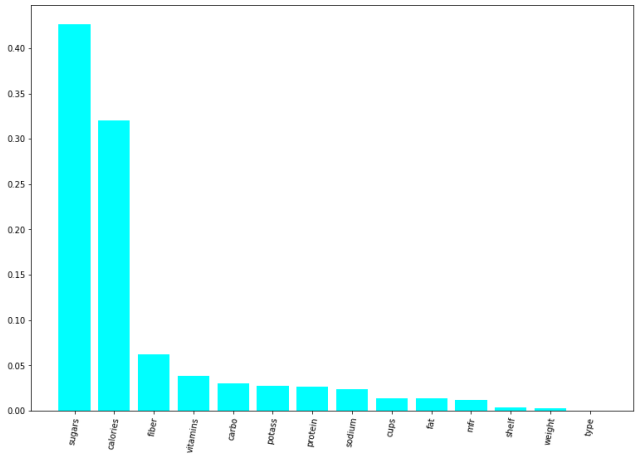

Figure 1: *Feature importances based on Random Forest Regression*

Pearson correlation heatmap;

We used Pearson correlation heatmap (*Figure 2*) see the correlation of independent variables with each other. Also, heatmap told us, is the features positive or negative proportional correlated.

According the *Figure 2* cell value at intersection of "type" and "rating" is nearly zero. It show us their correlation equal to 0. The strongest correlated intersection between ranking is "sugar" which is equal to -0.76. According this value, amount of sugar is the important feature for determine the rating, and less sugar means better rating.

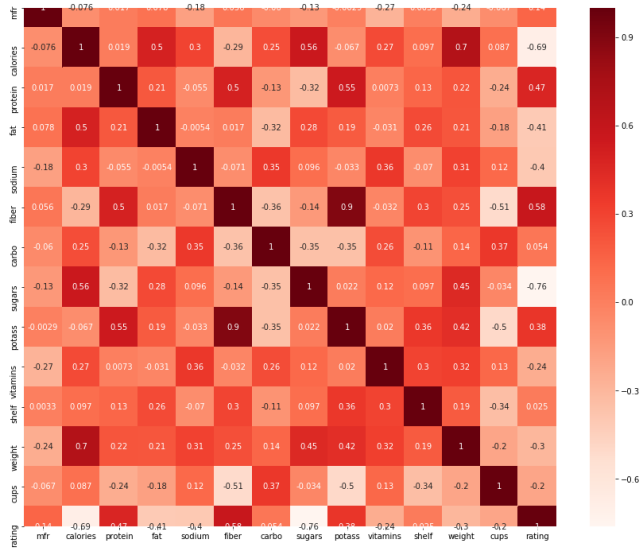

Figure 2: *Correlation heatmap*

D. Deletion

Less is more. First of all, we deleted the name of cereal feature, because It is all the values are unique. So they are meaningless for our models. Secondly, we decided to remove the type of cereal feature. Because, according to our feature importance analysis type data contributes almost zero.

IV. Methods & Results

A. Environment

We chose Python programming language for training and testing of our models with dataset. Python is a programming language, which used in data science because of algorithms and functions that used for data analysing and processing. As editor of Python, we use Anaconda and Spyder. They provide us lots of convenience. We can monetarize variable of data, output console and data frame. Spyder's variable explorer tool colourizing and sorting data frame values comprehensibly for us.

B. Machine Learning Methods

After preprocessing We choose one types of algorithm that could do regression predictions(for rating of cereal). We tried as many as possible, including but not

limited to:

- Linear Regression
- Support Vector Regression (SVR)
- Decision Trees
- Random Forests

Linear Regression;

We began to analyze our data by implementing linear regression first. Linear regression is a relatively simple algorithm where the algorithm finds a line in higher dimensions such that the sum of the squared distance between the line and the data points is minimized. After having fit this line, the algorithm predicts the outcome of an unseen data point by plugging in the point's features to the line equation.

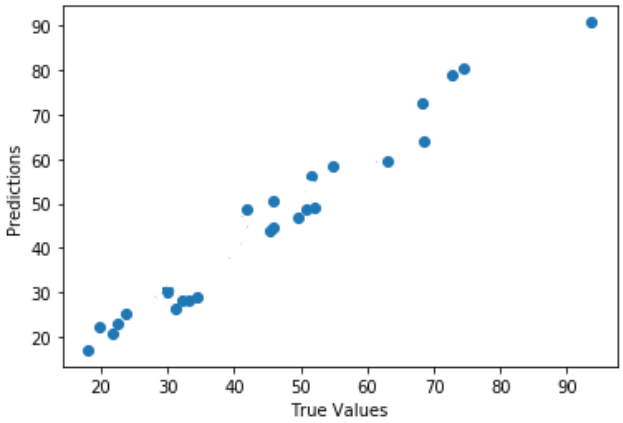

Figure 3: *Linear Regression Predictions*

We separated %30 of the samples for testing, and separated rest of them for training. Simple Linear Regression yielded %95.7 accuracy which is quite impressive. Model gave us 8.64916 mean squared error and 0.956728 R square error.

Support Vector Regression;

In simple regression we try to minimise the error rate. While in SVR we tried to fit the error within a certain threshold. We used it ,because for assuming the algorithm can reduce the error. [2]

Unfortunately Support Vector Regression algorithms effecting easily from outliers while creating models. So, we need to data standardization. We used "StandardScaler" is that it will transform our data such that its distribution will have a mean value 0 and standard deviation of 1.

Kernel Method	R Square	Mean Squared Error
RBF	0.91203	0.07960
Linear	0.97391	0.02360
Polinominal	0.78938	0.19058
Sigmoid	0.21933	0.70641

Table 1

We separated %20 of the samples for testing, and separate rest of them for training. We modeled 4 kernel methods which are the function used to map a lower dimensional data into a higher dimensional data for testing the results and use the best one.

According to different kernel methods, SVR model yielded different accuracy scores as we put it into *Table 1*. SVR yielded one of the best results among all the methods that we tried for all the different conditions that are applied. At its peak SVR with linear kernel resulted in a 97.3% accuracy.

Decision Trees;

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.[1]

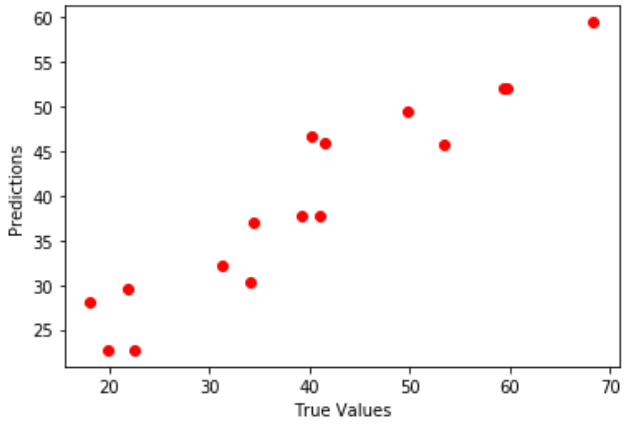

Figure 4: *Decision Trees Predictions*

Our decision tree regression model Performed %85.4 accuracy score with test data that is 20 percent of full samples. Mean squared error is 31.96502, and R square is 0.85447. It seems worse than other methods. But, R square results are not totally reliably score

measurement method especially for decision trees.

Random Forests;

Random Forests also employ Bagging, Bootstrap Aggregation. When a new tree is trained, it is trained using a uniformly random sample of features from the original dataset with replacement. Sampling with replacement allows for a sample of training features to be repeated. This is "Bootstrapping"[3].. Then, once all the trees are trained, they can be fed the testing data. Every tree will give a different result for the test data, but the average for regression.

Random Forests have many attractive qualities. Because Random Forests are essentially a collection of Decision Trees, they don't need to be trained with normalized data, and can handle both continuous and categorical data simultaneously. Much like decision trees, they have white box qualities, as you can look at the trees that produce the results. One can also calculate the importance of variables by averaging the error difference as the values of a variable are permuted across all the trees. Bagging also makes Random Forests resilient against variance and overfitting. It's likely that those are the reasons why it performed so well with our data.

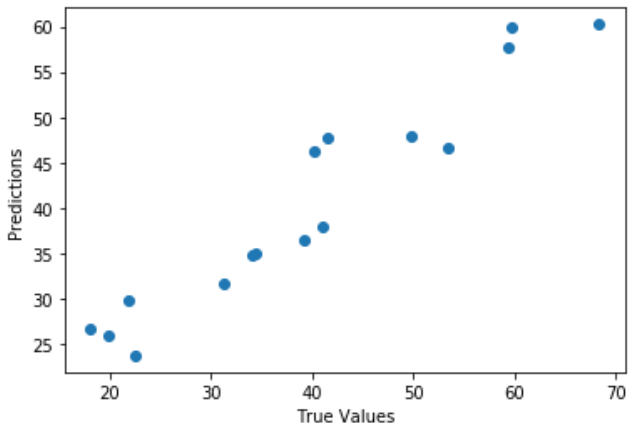

Figure 5: *Decision Trees Predictions*

In our model, test sample randomly chosen from %20 of data samples. According to comparison between true Values and predicted values %89.0 accuracy be monitored(*Figure 5*) with 3.89225 mean absolute error.

V. Conclusions & Discussion

Algorithms	Accuracy
Linear Regression	95.7%
Support Vector Regression	**97.4%**
Decision Trees	85.4%
Random Forests	89.0%

Table 2: *Results*

We used accuracy, precision and recall on the training and validation sets to evaluate the performance of each algorithm (Table 2).

Although all algorithms have considerable predictions, The Decision Trees algorithm gave a the least accuracy of 85.4%, with predictions on data points predict poorly to actual customer ratings.

Although LR and SVR yielded a very good accuracy of 95.7% and 97.7% against the validation data, with similar accuracy against the training data indicating no overfitting, SVR with Linear Kernel method is the "chef's recommendation".

For the SVR, each kernel did not yield reasonable accuracy on the test data. Especially, sigmoid method have poor accuracy on the validation data.

Using random forests did not improve the precision or recall. We could potentially increase the precision by collating a larger validation set.

Our features have a smooth, nearly linear dependence on the covariates, then linear regression will model the dependence better than random forests, which will basically approximate a linear curve with an ugly irregular step function. If the dependence is multivariate linear and smooth, with significant covariates producing the dependence, the fit performance of random forests can be expected to get worse and worse for larger and larger. RF has a much greater ability than a single decision tree to model linearity, since we are adding tree predictions together - but still, it's just not very efficient to approximate a high-dimensional linear relationship with a series of step functions. We think this is the most likely theoretical explanation for RF underperforming linear regression.

References

[1] Prashant Gupta(2017). Decision Trees in Machine Learning *https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052*

[2] M. Limam, W. Gani and Taleb, (2010). Vector regression based residual control charts *Journal of Applied Statistics*

[3] Breiman, Leo (1996). Bagging predictors *Machine Learning*.24 (2): 123–140.