Yield Prediction using Machine Learning and Data Analytics

Introduction

Prediction attempts to form patterns that permit it to predict the next event(s) given the available input data. Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value. Here our objective is to develop a model to predict percent yield of the manufacturing process.

Dataset Description

This data set contains information about a chemical manufacturing process, in which the goal is to understand the relationship between the process and the resulting final product yield. The data set consisted of 177 samples of biological material for which 57 characteristics were measured. Of the 57 characteristics, there were 12 measurements of the biological starting material, and 45 measurements of the manufacturing process. The process variables included measurements such as temperature, drying time, washing time, and concentrations of by–products at various steps. Some of the process measurements can be controlled, while others are observed.

Predictors are continuous, count, categorical; some are correlated, and some contain missing values. Samples are not independent because sets of samples come from the same batch of biological starting material.

Data Preprocessing

Learning algorithms have affinity towards certain data types on which they perform incredibly well. They are also known to give reckless predictions with unscaled or unstandardized features. In simple words, pre-processing refers to the transformations applied to your data before feeding it to the algorithm.

Here a data frame with columns for the outcome (Yield) and the predictors (BiologicalMaterial01 though BiologicalMaterial12 and ManufacturingProcess01 though ManufacturingProcess45).

Biological predictors cannot be changed but can be used to assess the quality of the raw material before processing. On the other hand, manufacturing process predictors can be changed in the manufacturing process. Improving product yield by 1 % will boost revenue by approximately
one hundred thousand dollars per batch.

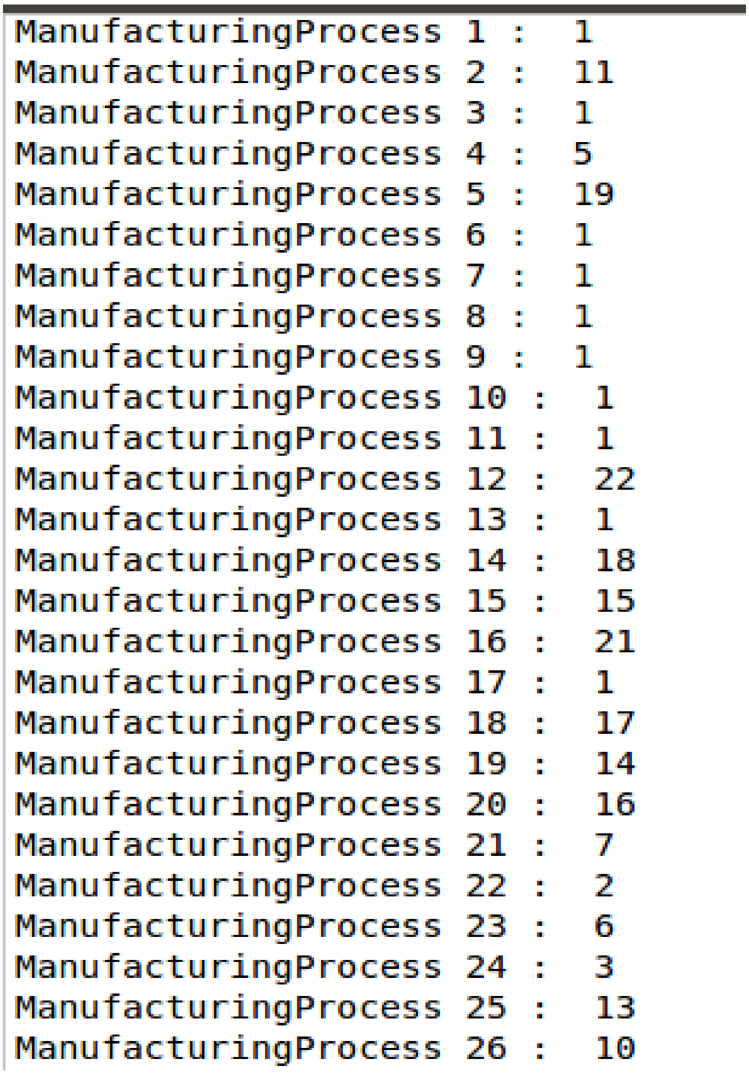
1.1) Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. Feature selection methods can be used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model.

There are different methods for performing feature selection on dataset:
1) Correlation
2) Forward selection
3) Recursive feature elimination
4) Feature importance

Here we used Recursive feature elimination for removing the irrelevant attributes. The Recursive Feature Elimination (or RFE) works by recursively removing attributes and building a model on those attributes that remain. It uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

We perform feature selection only on “ManufacturingProcess” attributes(on 45 attributes) because “biological” predictors cannot be changed. After applying RFE it will assign rank with each feature according to their importance.



```
ManufacturingProcess 1 :  1
ManufacturingProcess 2 :  11
ManufacturingProcess 3 :  1
ManufacturingProcess 4 :  5
ManufacturingProcess 5 :  19
ManufacturingProcess 6 :  1
ManufacturingProcess 7 :  1
ManufacturingProcess 8 :  1
ManufacturingProcess 9 :  1
ManufacturingProcess 10 :  1
ManufacturingProcess 11 :  1
ManufacturingProcess 12 :  22
ManufacturingProcess 13 :  1
ManufacturingProcess 14 :  18
ManufacturingProcess 15 :  15
ManufacturingProcess 16 :  21
ManufacturingProcess 17 :  1
ManufacturingProcess 18 :  17
ManufacturingProcess 19 :  14
ManufacturingProcess 20 :  16
ManufacturingProcess 21 :  7
ManufacturingProcess 22 :  2
ManufacturingProcess 23 :  6
ManufacturingProcess 24 :  3
ManufacturingProcess 25 :  13
ManufacturingProcess 26 :  10
```

Out of those 45 attribues, we selected 24 attributes based on their correlation value with output variable.

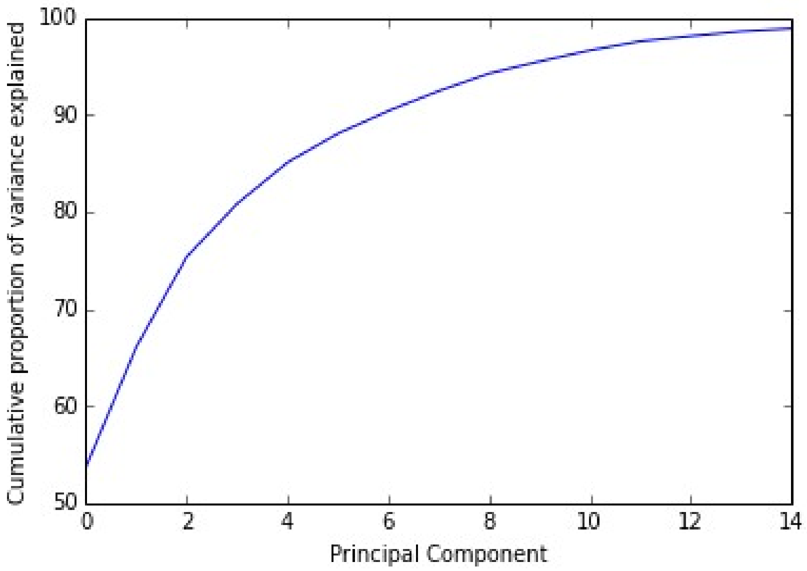
1.2) Dimension Reduction

Dimension Reduction refers to the process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely. These techniques are typically used while solving **machine learning problems** to obtain better features for a classification or regression task.

For example, let us take case of a motorbike rider in racing competitions. Today, his position and movement gets measured by GPS sensor on bike, gyro meters, multiple video feeds and his smart watch. Because of respective errors in recording, the data would not be exactly same. However, there is very little incremental information on position gained from putting these additional sources. Now assume that an analyst sits with all this data to analyze the racing strategy of the biker – he/ she would have a lot of variables / dimensions which are similar and of little (or no) incremental value. This is the problem of high unwanted dimensions and needs a treatment of dimension reduction.

Here we used **Principal Component Analysis** (PCA) for dimension reduction. In this technique, variables are transformed into a new set of variables, which are linear combination of original variables. These new set of variables are known as **principle components.** They are obtained in such a way that first principle component accounts for most of the possible variation of original data after which each succeeding component has the highest possible variance.

Here we applied PCA on the attributes which we got from feature selection (24 attributes).



This plot shows that 15 components results in variance close to ~ 98%. Therefore, in this case, we'll select number of components as 15 [PC1 to PC15] and proceed to the modeling stage.

Modeling

Our problem here is to predict the yield based on some information of input vaiables. Here our output variable (yield) contain continuous value and our problem is prediction problem. In Machine Learning, there is subclass of prediction which contain the algorithms that can be used for prediction problem.

Mostly for prediction problem when output variable is continuous, Regression algorithm were used. Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variables (predictor). This technique is used for forecasting, predictive modelling and finding the causal effect reltionship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Here we used Linear Regression (LR) and Bayesian Ridge Regression (BRR). First we checked the output without applying anything on dataset, model is build using linear regression algorithm. Results of applying different models on dataset are shown below.

Performance measure \ models	**Linear regression (LR)**	**RFE+LR**	**RFE+PCA+ LR**	**RFE+BRR**
Correlation	0.74	0.82	0.82	0.83
RMSE	1.3907	1.051	1.020	0.998

Where
RFE- Recursive Feature Elimination
LR- Linear Regression
PCA- Principal Component Analysis
BRR- (Bayesian Ridge Regression)
RMSE- Root Mean Square Error

Performance measures:

1) RMSE (Root Mean Square Error)

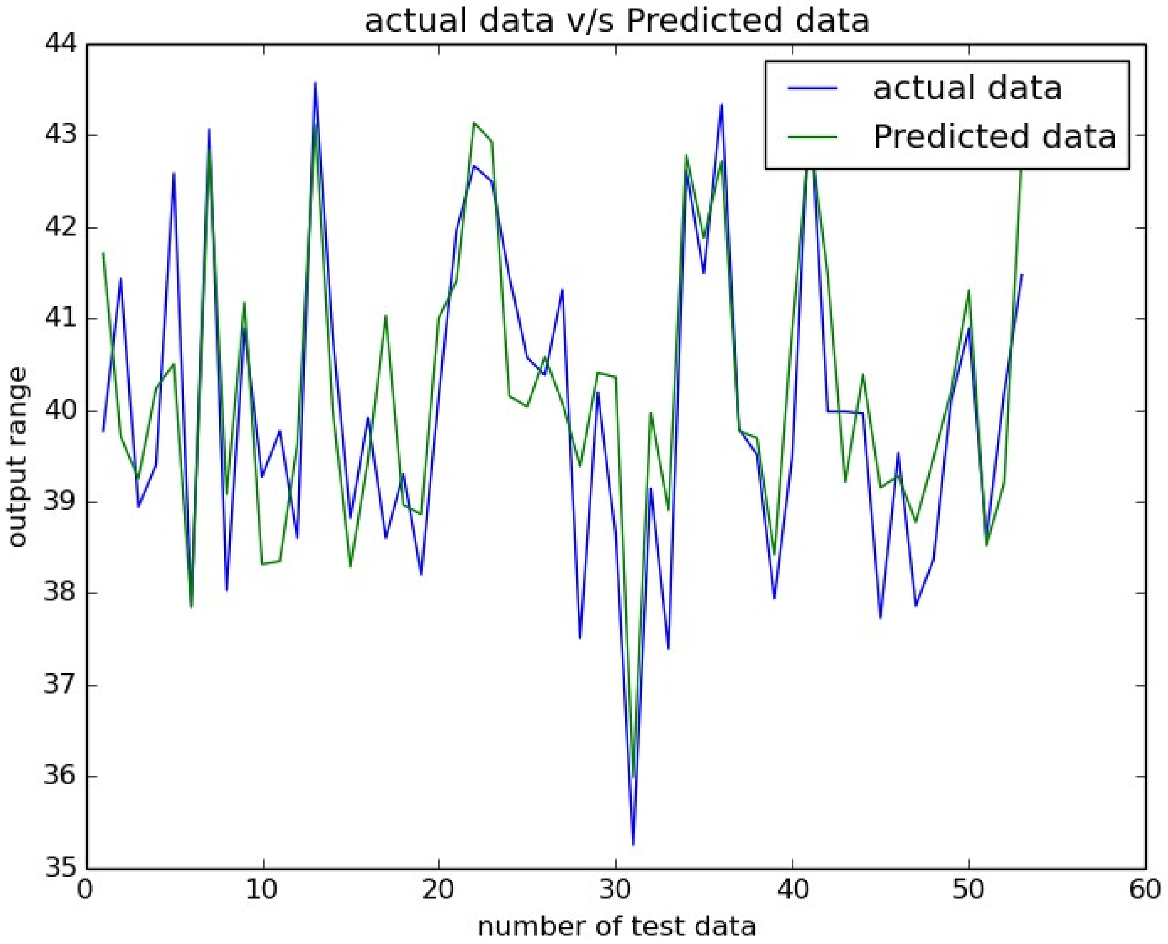
The RMSE is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. The RMSD represents the sample standard deviation of the differences between predicted values and observed values.

2) Correlation cofficient

A correlation coefficient is a number that quantifies some type of correlation and dependence, meaning statistical relationships between two or more random variables or observed data values. It's value is between [-1,+1].

-1 -> Perfect negative linear correlation
+1 -> Perfect positive linear correlation
0 -> no correlation

Our goal is to increase the correlation and reduce the RMSE value. From the result shown above, first we build model using Linear regression without doing Feature Selection and Dimension Reduction. We got improvement in result after doing feture selection. Like this we improved out result with builing the model using Bayesin Ridge Regression (RMSE:0.998 and correlation:0.83).



This graph show the relation between the actual value and predicted value, when we apply Bayesian Ridge Regression with RFE.