DARIA KUDOIAR
dariakudoiar@gmail.com

Introduction:

For the formation of investment strategies and the development of risk management models, Stock price prediction mechanisms are Fundamentally important. However, it is not really possible to consistently obtain risk-adjusted returns above the profitability of the market as a whole. Computational advances have led to several machine learning algorithms used to anticipate market movements consistently and thus estimate future asset values such as company stock prices. Models based on the Kernel Ridge Regression (KRR) are among the most widely used techniques.

Question 1 Exploring Data Set

In this project, we will try to predict the daily closing stock price of Goldman Sachs Group Inc. (GS). To better understand the price movements, we visualize the stock prices for GS over ~ 9 years. From the plot we can see significant variation of the prices. The stock is not cyclical but is rather driven by external factors. As an investment bank, Goldman Sachs depends on the global economy**.** Bad or volatile economy possibly limited proprietary trading earnings.

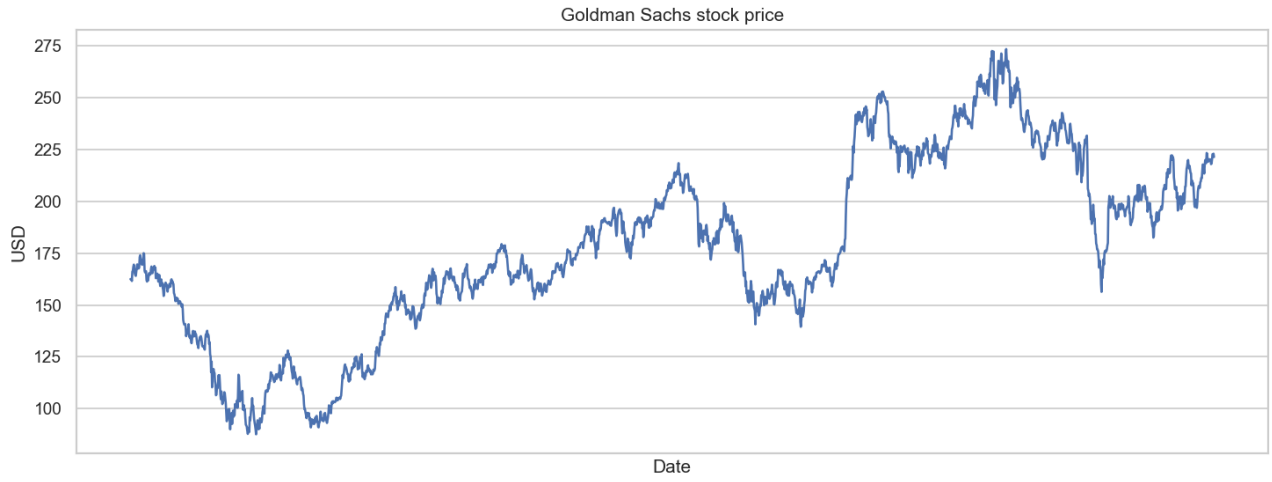


Figure 1: Goldman Sachs Group stock prices for the period Dec, 2010 to Nov, 2019

To predict the daily closing price, we will use historical data from Yahoo! Finance for 11 Banks, such as: Bank of America Corp., Bank of New York Mellon Corp., Capital One Financial Corp., Citigroup Inc., Wells Fargo & Co., JPMorgan

Chase & Co., Morgan Stanley, PNC Financial Services Group Inc., TD Group US Holdings LLC, U.S. Bancorp and Goldman Sachs Group Inc. We will use daily closing prices from December 2nd, 2010 to November 29th, 2019 (approximately nine years). To achieve the task of prediction we will use Kernel Ridge Regression method, which is a penalized regression that uses the kernel trick. Kernel is now being used in a lot of machine learning algorithms. Basically, it transports the data to a higher hyper plane where it almost becomes linear.

The main technical goal is to apply Kernel Ridge Regression (KRR) to predict the value of stock prices for Goldman Sachs Group Inc whenever a new case X = [X1... Xp] is given. However, accurately predicting stock price movements is an extremely complex task, so the more we know about the stock (from different perspectives) the higher our changes are. There are millions of events and pre-conditions for a particular stock to move in a particular direction.

The direction of the stock market index refers to the movement of the price index or the trend of fluctuation in the stock market index in the future. Predicting the direction is a practical issue that heavily influences a financial trader's decision to buy or sell an instrument. Accurate forecast of the trends of the stock index can help investors to acquire opportunities for gaining profit in the stock exchange.

In machine learning algorithms, what we do is learn influences to the past performance to predict the future. However, the influences of the past are very small and stock prices are often more oriented to the future performance of a company and its sector. Also, it seems that quality of outcome is depended on various factors, if we able to accommodate all the factors into learning system, the result should improve.

Original data set consists of 2265 cases which represent 2265 days of historical data for each bank. We take a sample of original dataset in regard to reduce computational time. We get a data set with a shape of 1000 cases and 12 columns, such as: Date, 10 explanatory variables (x1, … x10) and an observed value y.

Date (December 2nd, 2010 to November 29th, 2019)

x1 = Bank of America Corp.

x2 = Bank of New York Mellon Corp.

DARIA KUDOIAR
dariakudoiar@gmail.com

x3 = Capital One Financial Corp.

x4 = Citigroup Inc.

x5 = Wells Fargo & Co.

x6 = JPMorgan Chase & Co.

x7 = Morgan Stanley

x8 = PNC Financial Services Group Inc.

x9 = TD Group US Holdings LLC

x10 = U.S. Bancorp

y = Goldman Sachs Group Inc.

We include an artificial feature $X_p(j) = 1$ for all cases $j=1...n$.

We have all features as continuous variables. We calculate mean and standard deviation for each feature and response variable.

Features x_i and response variable y	***MEAN***	***STANDARD DEVIATION***
x1 Bank of America Corp.	18.03	7.52
x2 Bank of New York Mellon Corp.	38.4	10.79
x3 Capital One Financial Corp.	73.25	16.08
x4 Citigroup Inc.	51.99	13.17
x5 Wells Fargo & Co.	45.91	9.89
x6 JPMorgan Chase & Co.	69.54	26.49
x7 Morgan Stanley	33.45	11.35
x8 PNC Financial Services Group Inc.	94.76	30.63
x9	55.6	13.26

TD Group US Holdings LLC		
x10 U.S. Bancorp	41.89	9.4
y Goldman Sachs Group Inc.	177.97	43.11

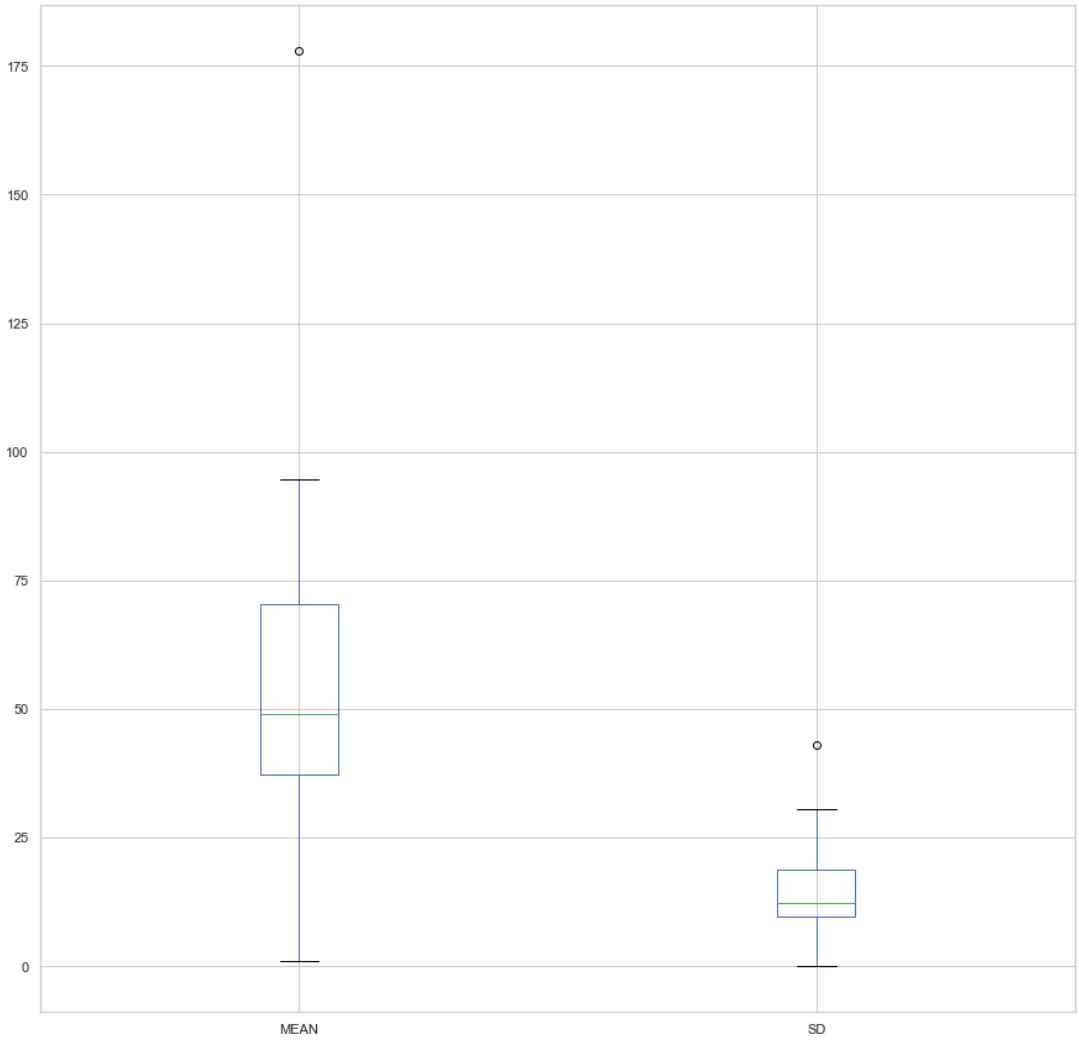


Figure 2: Mean and Standard Deviation of features

The box plots above show the spread of the means and standard deviations of the features. Boxplot for means is slightly skewed right and it may indicate bigger variation among the mean of features somewhere between 50 to 70. In addition, boxplot for means indicates “outlier” with value ~180.

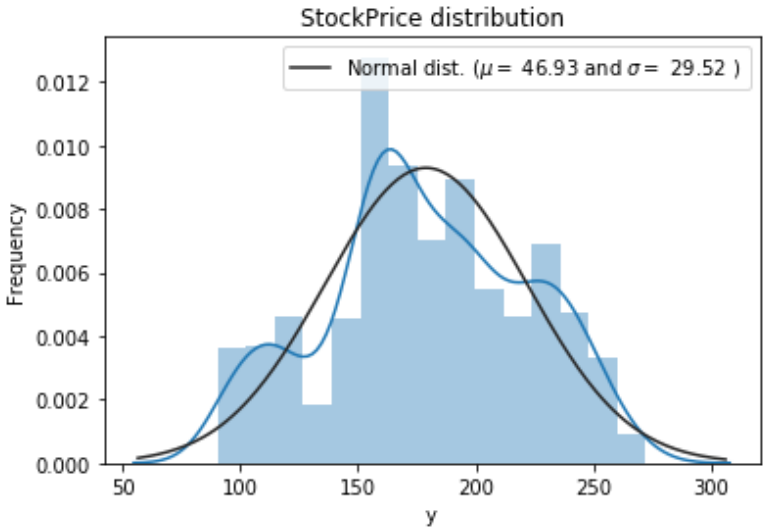
DARIA KUDOIAR
dariakudoiar@gmail.com

Boxplot for standard deviations are somewhat symmetrical. Again, we see outlier for value ~ 45.

For further analysis we set index in data frame as “Date” and split data set into train and test set with respective proportions 80% and 20%. The breakdown of the numbers of cases in the training and test data sets is:

X train	800 cases, 11 columns
X test	200 cases, 1 column
Y train	800 cases, 11 columns
Y test	200 cases, 1 column

In order to make sure our data is suitable we want to perform a couple tests to analyze target variable to ensure that the results we achieve and observe are indeed real, rather than compromised due to the fact that the underlying data distribution suffers from fundamental errors. Since ensuring that the data has good quality is very important for our models. We check distribution for response variable as well as Probability Plot.



DARIA KUDOIAR
dariakudoiar@gmail.com

Figure 3: Distribution of response variable

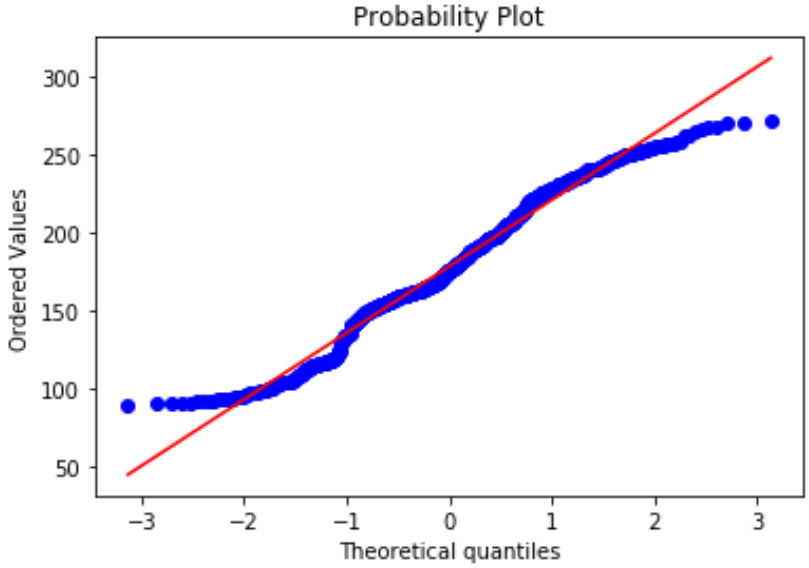


Figure 4: Normal Q-Q plot for response variable

The first thing that can be observed is that points form a straight line rather than a curve, which means there is no serious indication of skewness in the sample data. By looking at the tails of the distribution, we see that left and right tails are slightly heavier than we would like them to be in 'ideal' situation. However, we can assume that the target variable doesn't seem to show any significant problem. A linear pattern in the points indicates that the given family of distributions reasonably describes the empirical data distribution. As regression models perform better with normally distributed data, we would like to make sure variable is normally distributed.

We compute empirical correlations cor(x_i, y) ... cor(x_p, y) and their absolute values C_1 ... C_p. We create a feature correlation heatmap to visualize the result.

The colors on the heatmap represent how strong correlation between explanatory variables and response variable. The highest correlation has values that close to 1 or – 1, we should look at the lightest and the most intense colors on the heatmap. We see that the lowest correlation between features x5, x9 and y, however the correlation is still quite high, 0.85. The strongest correlation between x7 and y, which is 0.95.

DARIA KUDOIAR
dariakudoiar@gmail.com

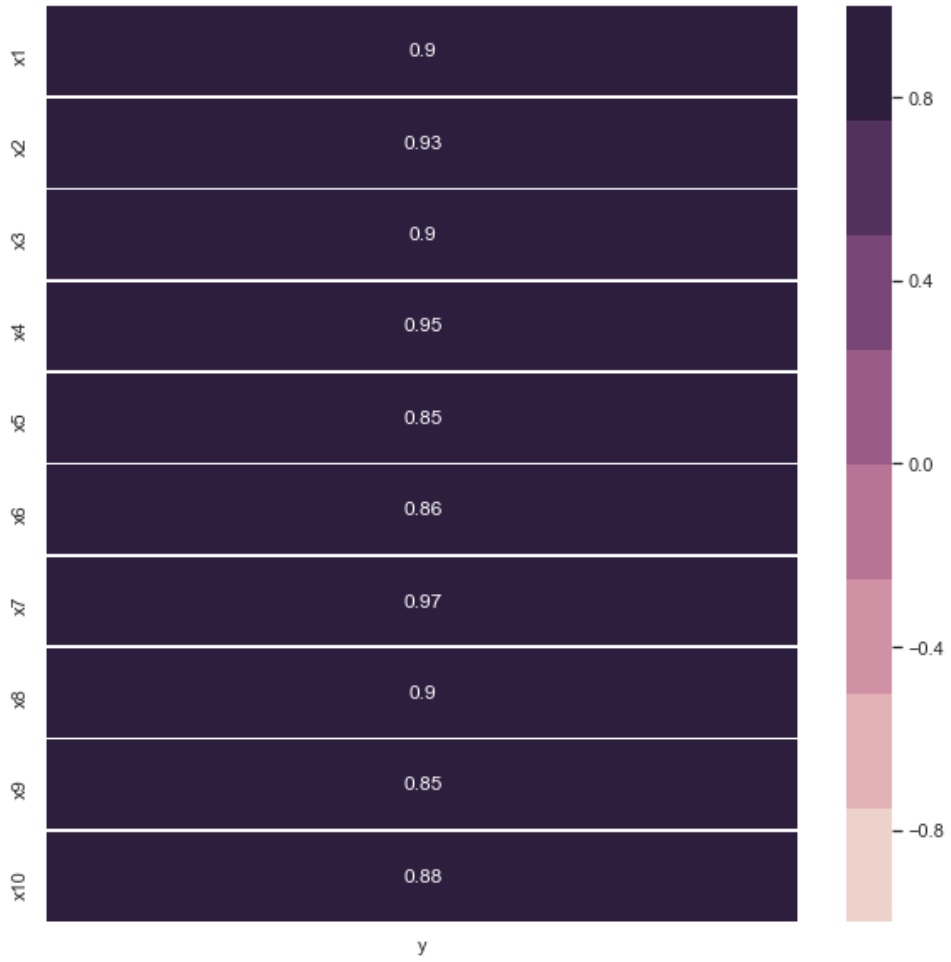


Figure 5: Feature Correlation Heatmap

We compute the 3 largest values among C_1 ... C_p, which are C_u(x7, y) = 0.97, C_v(x4, y) = 0.95 and C_w(x2, y) = 0.93. And we display three scatterplots with the highest correlation features vs target variable.

DARIA KUDOIAR
dariakudoiar@gmail.com

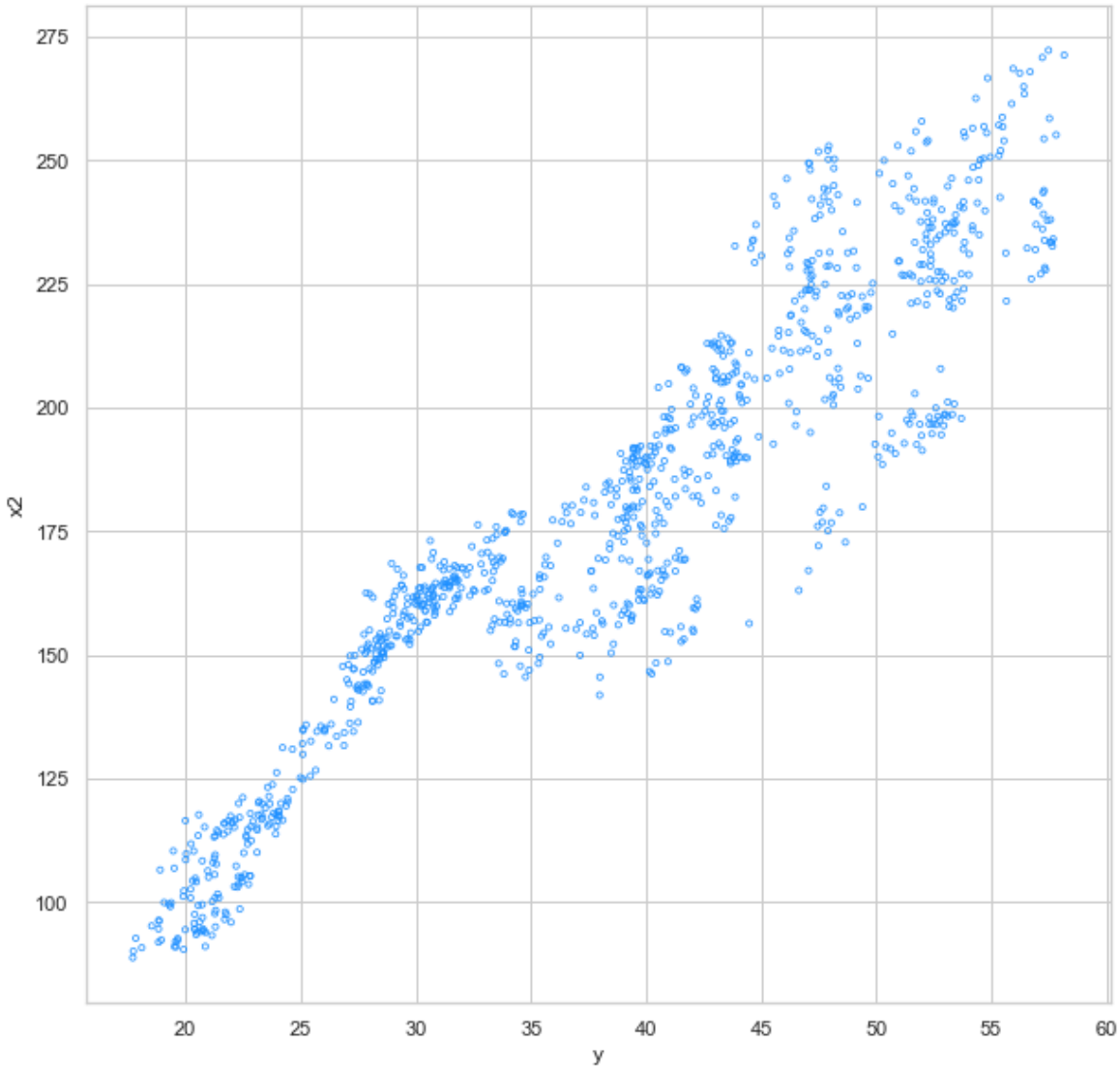


Figure 6: Scatter plot for correlation between feature x2 and response variable y (Bank of New York Mellon Corp and Goldman Sachs Group Inc.)

From scatterplot we see somewhat strong (but not perfect) positive correlation between feature x2 and response variable, however the data points highly scattered between 33 – 60 for y values. The data has a large variance, but x2 could be a helpful feature to predict stock price of y.

DARIA KUDOIAR
dariakudoiar@gmail.com

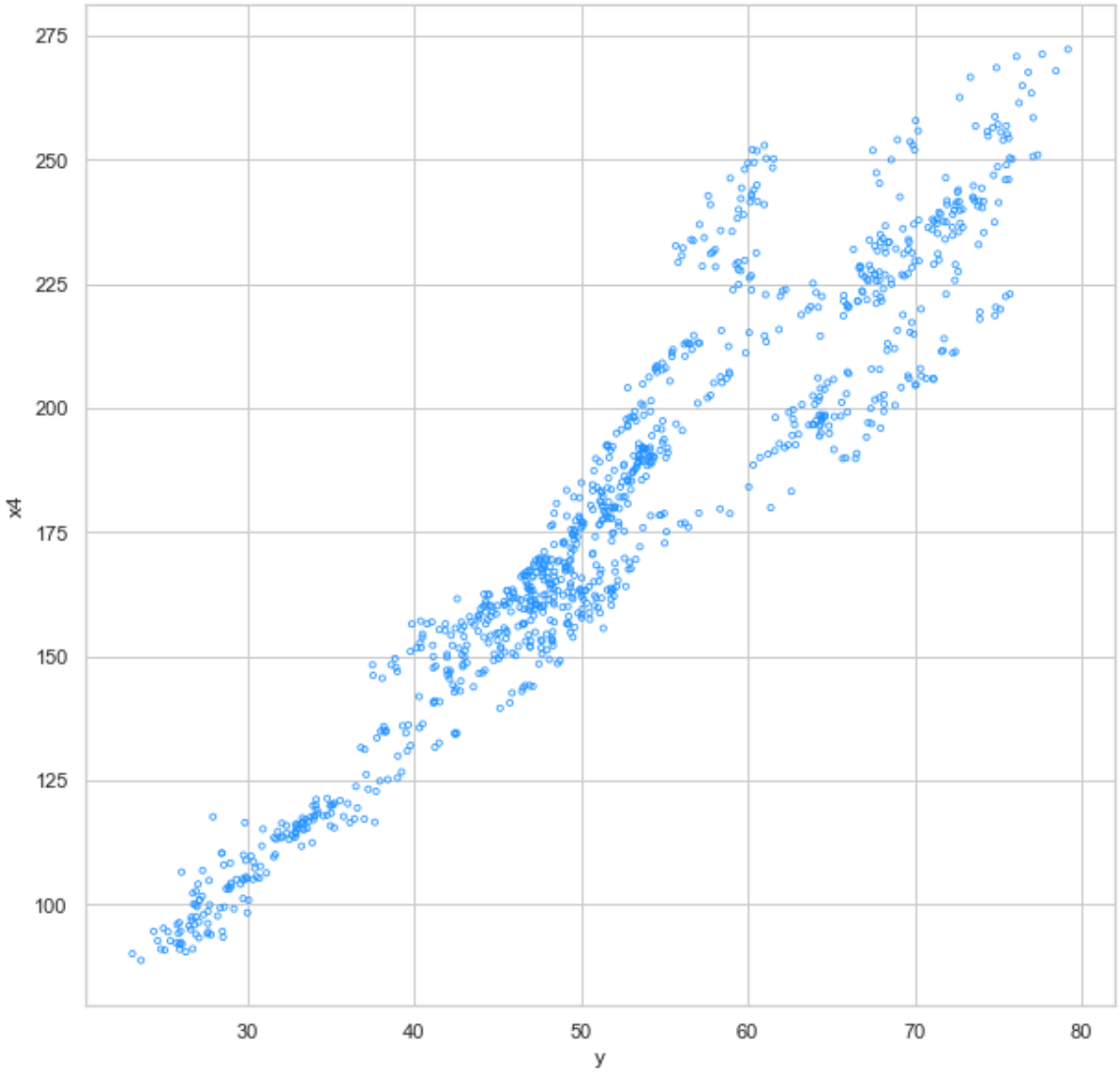


Figure 7: Scatter plot for correlation between feature x4 and response variable y (Citigroup Inc and Goldman Sachs Group Inc.)

The scatterplot for feature x4 in relation to response variable y, indicates a strong positive correlation. Data points are grouped mostly close together with similar spreads for each y value. However, scatter plot displays observations that are closely grouped between 10 to 55 and spreads out more when you increase variable y. This feature may play a good role in predicting the dependent variable.

DARIA KUDOIAR
dariakudoiar@gmail.com

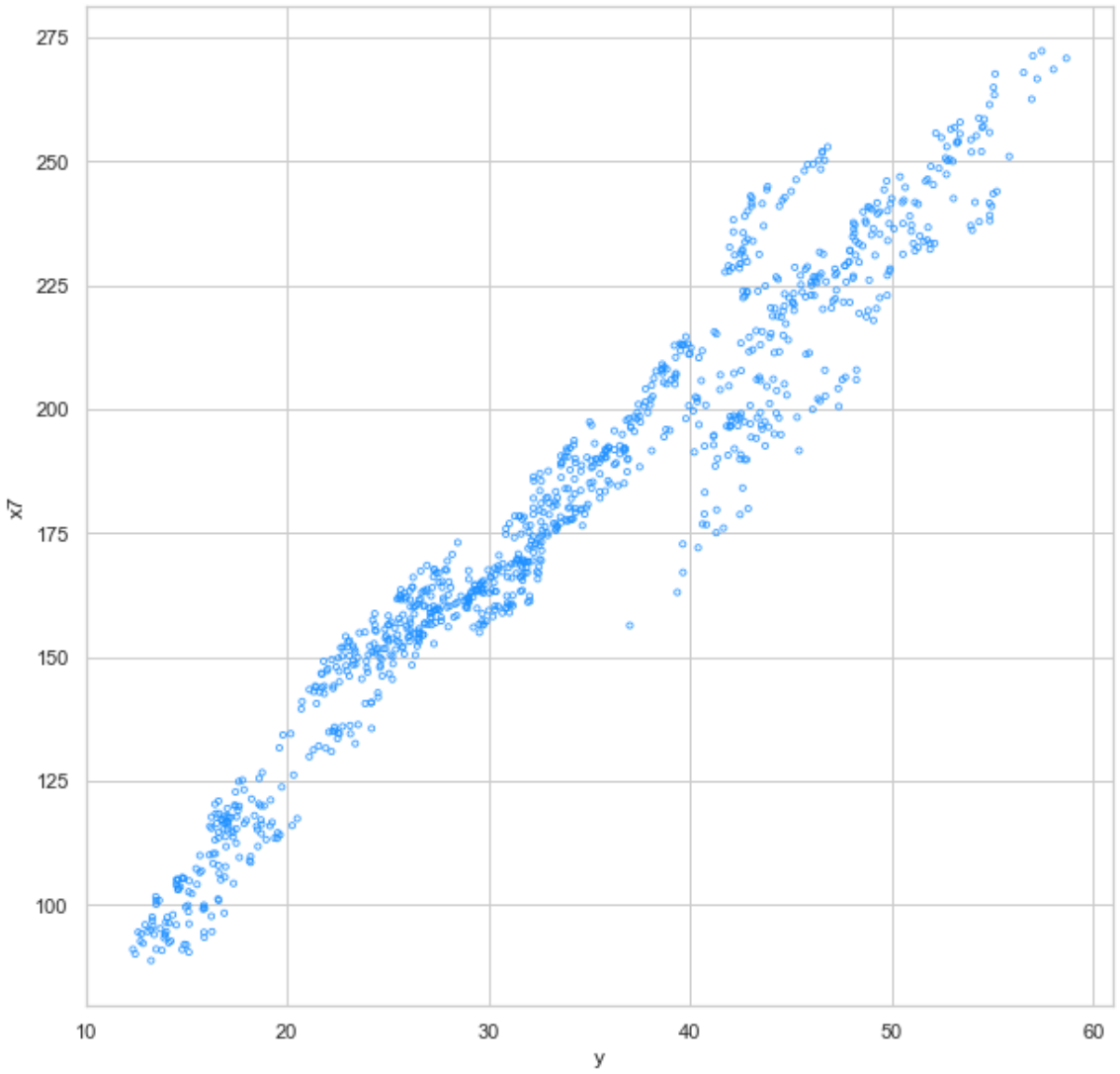


Figure 8: Scatter plot for correlation between feature x7 and response variable y (Morgan Stanley and Goldman Sachs Group Inc.)

Compared to the previous scatterplots we can conclude that scatterplot for y vs x7 indicates the strongest positive correlation. Data points are clearly arranged in a linear fashion, however data points spread out more somewhere between 40 ~ 50 for y value, but still the data is concentrated together to form a linear pattern. Feature x7 could play significant role in predicting y.

DARIA KUDOIAR
dariakudoiar@gmail.com

QUESTION 2 Kernel Ridge Regression (KRR) with radial kernel

We use the training set for further calculations, size of the train set is 800 cases. To select parameters λ and gamma we randomly choose two lists of 100 numbers within train set [1...m] and for all i in List1 and all j in List2 we compute distances given by the formula:

$$D_{ij} = \|X(i) - X(j)\|^2$$

We plot the histogram of the 10000 numbers D_{ij} and compute q =10% quantile of the 10000 numbers D_{ij} and set gamma = 1/q = 0.073

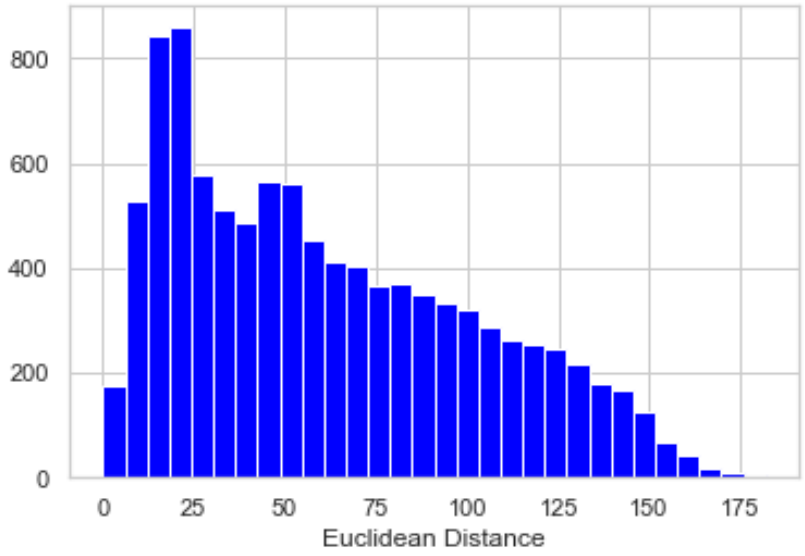


Figure 9: Histogram of the 10000 numbers D_{ij}

We use fixed gamma to calculate Gramian (mxm matrix) which is the Kernel Gramian:

$$G = [G_{ij}]$$

Where $G_{ij} = K(X(i), X(j))$ for all i, j in [1...m]. We compute the matrix G using Radial or RBF Kernel, given by the formula:

$$K(x, y) = e^{-\gamma\|x-y\|^2}$$

DARIA KUDOIAR
dariakudoiar@gmail.com

and eigenvalues of G, L1 > L2 > ... > Lm ≥ 0. We plot decreasing curve for eigenvalues versus j, where j = 1… 800 as well as calculate and plot increasing ratios R_j= (L1 + ... + Lj)/(L1+ ... + Lm).

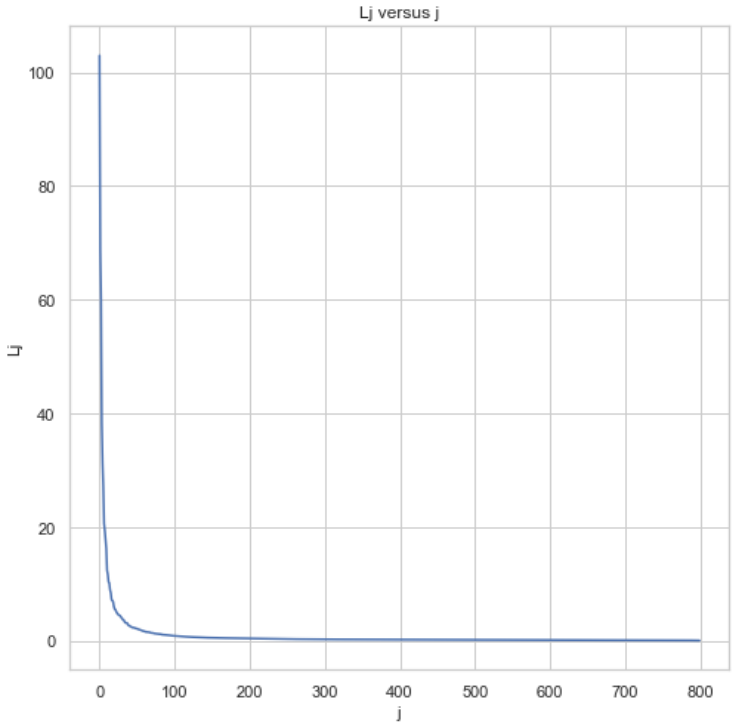


Figure 10: Plot Lj versus j

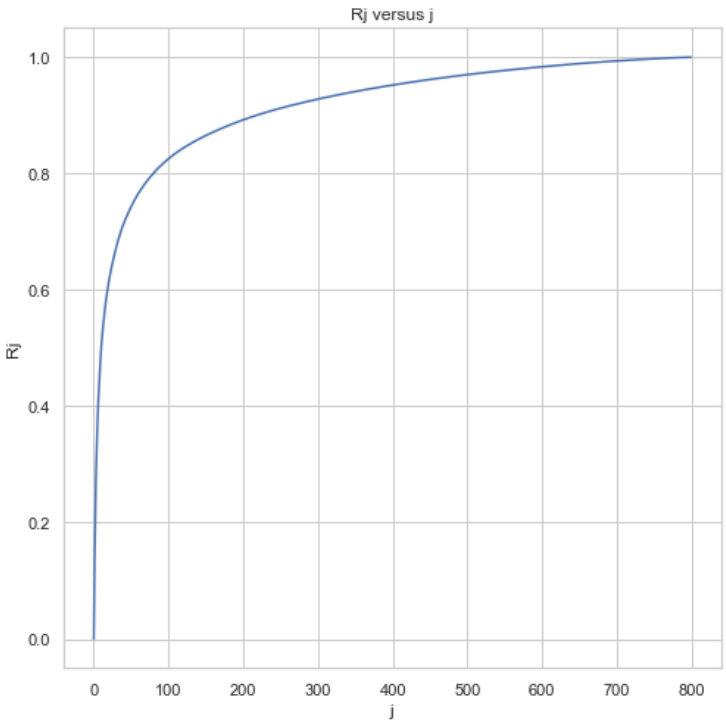


Figure 11: Plot Rj versus j

The scatter plot *Figure 10* represents the eigenvalues for each of the 800 predictors. The curve rapidly decreases, and around λ20, we notice the "elbow". After the "elbow", the λj values decrease gradually. The scatterplot *Figure 11* represents the ratios of eigenvalues. The smallest j = 378 such that RATj $\geq$ 95% and fix λ = 0.17.

Now we ready to calculate matrix M = G + λ Id and its inverse M^{-1}. Where Id is (m×m) identity matrix, G is Gramian and $\lambda > 0$ is a parameter that roughly evaluates the cost of a prediction error.

Once λ and gamma are selected, the best KRR prediction function pred(x) is defined for any input vector x in R^p by the formula:

$$pred(x) = y\,(G + \lambda Id)^{-1}V(x)$$

DARIA KUDOIAR
dariakudoiar@gmail.com

y= [Y1 ... Ym] is a line vector, V(x) is a column vector with m coordinates

$V_1(x)$, ... , $V_m(x)$ given by $V_j(x) = K(x, X(j))$

Prediction formula becomes:

$$pred(x) = A1\ K(x, X(1) + \cdots + Am\ K(x, X(m)$$

where A is a line vector [A1 ... Am] given by formula by $A = yM^{-1}$,

We calculate the prediction and calculate mean square error (MSE), root squared error (RMSE) and ratios of RMSE/average y. RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. RMSE ratio indicates the robustness of learning. Ideally, we want to have both RMSE and RMSE ratio as small as possible.

Data	*RMSE*	*RMSE ratio*	*95% Confidence interval for RMSE ratio*
Train set	17.5	0.098	[0.03 ; 0.166]
Test set	53.4	0.307	[0.18 ; 0.68]

We get train RMSE and test RMSE of 17.5 and 53.4, respectively, which by itself is not quite good result. We expect to have higher mean square errors on Test set as we see from our results, because the supposed patterns that the method found in the training data don't really exist in the test data.

Since the confidence intervals do not overlap, we can say there is significant difference in the Root Mean Square Error between the training data and the test data.

We also want to compare predicted values with original values of y. We display scatterplot of y original vs y predicted of Test set. As we can see from the scatterplot the model gives not very good approximation of the real stock price.

DARIA KUDOIAR
dariakudoiar@gmail.com

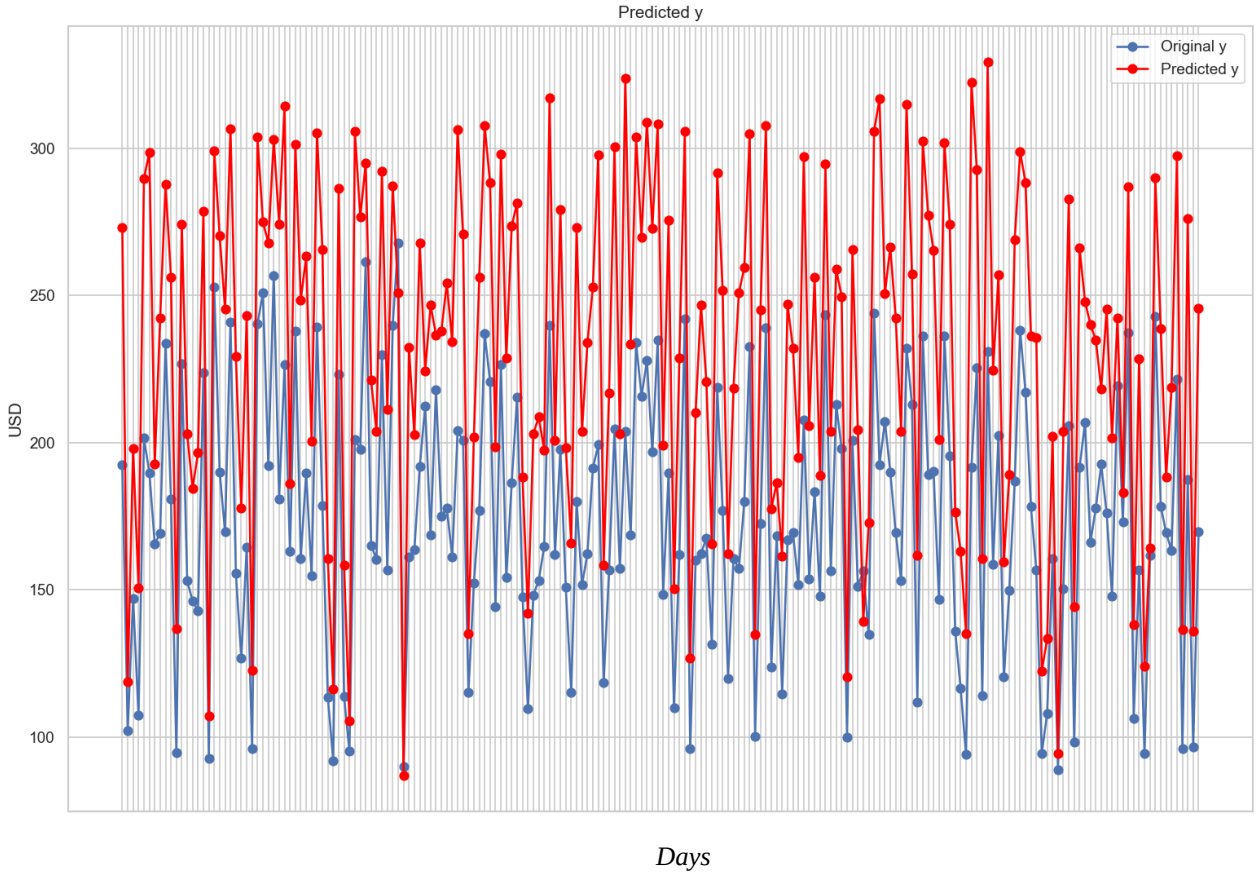


Figure 12: Original stock prices vs Predicted stock prices for Goldman Sachs Group on Test set

Scatterplot indicates somewhat good results for the first model with defined prediction formula and fixed parameters gamma and λ. For some points we clearly see that there is very small difference between two points true value of y and predicted value of y, which means for those cases RMSE is very low.

DARIA KUDOIAR
dariakudoiar@gmail.com

QUESTION 3

To improve the results of prediction we tune parameters through step by step tuning. We repeat the whole procedure for Q2 with different pairs of parameters gamma and λ. We change one parameter at a time and decide about direction of improving performances.

Parameters		RMSE		RMSE ratio	
Gamma	λ	Train	Test	Train	Test
0.001	0.17	330.17	1.85	332.63	1.91
0.01	0.17	165.09	0.92	164.53	0.94
0.1	0.17	135.95	0.76	131.36	0.75
1	0.17	4456.82	24.91	4263.85	24.48
0.01	0.9	161.46	0.9	159.26	0.91
0.069	1.52	9.07	0.05	11.36	0.07
0.06	2	37.68	0.21	37.97	0.22
0.073	1	11.43	0.06	12.12	0.07
0.073	1.5	11	0.06	11.94	0.07
0.072	1.5	8.46	0.05	10.09	0.06
0.07	1.5	7.23	0.04	9.76	0.06
0.07	1	5.51	0.03	8.23	0.05
0.07	0.95	5.32	0.03	8.08	0.05
0.07	0.9	5.13	0.02	7.91	0.04

From the table we see, that both situations with decreasing and increasing Gamma and λ = const, performance on Train and Test set shows the very high RMSE and the ratio RMSE/av_y. We have slightly better situation with smaller gamma. We fix gamma and try to increase λ, we get the better results for RMSE.

We select the best choice of parameters in terms of accuracy RMSE/avy and stability of performance when one goes from TRAIN to TEST set. The best parameters are: gamma = 0.07 and λ = 0.9. We compute the prediction for y with the best parameters and get the following results.

Data	*RMSE*	*RMSE ratio*	*95% Confidence interval for RMSE ratio*
Train set	5.13	0.03	[0.02; 0.04]
Test set	7.91	0.05	[0.02; 0.08]

DARIA KUDOIAR
dariakudoiar@gmail.com

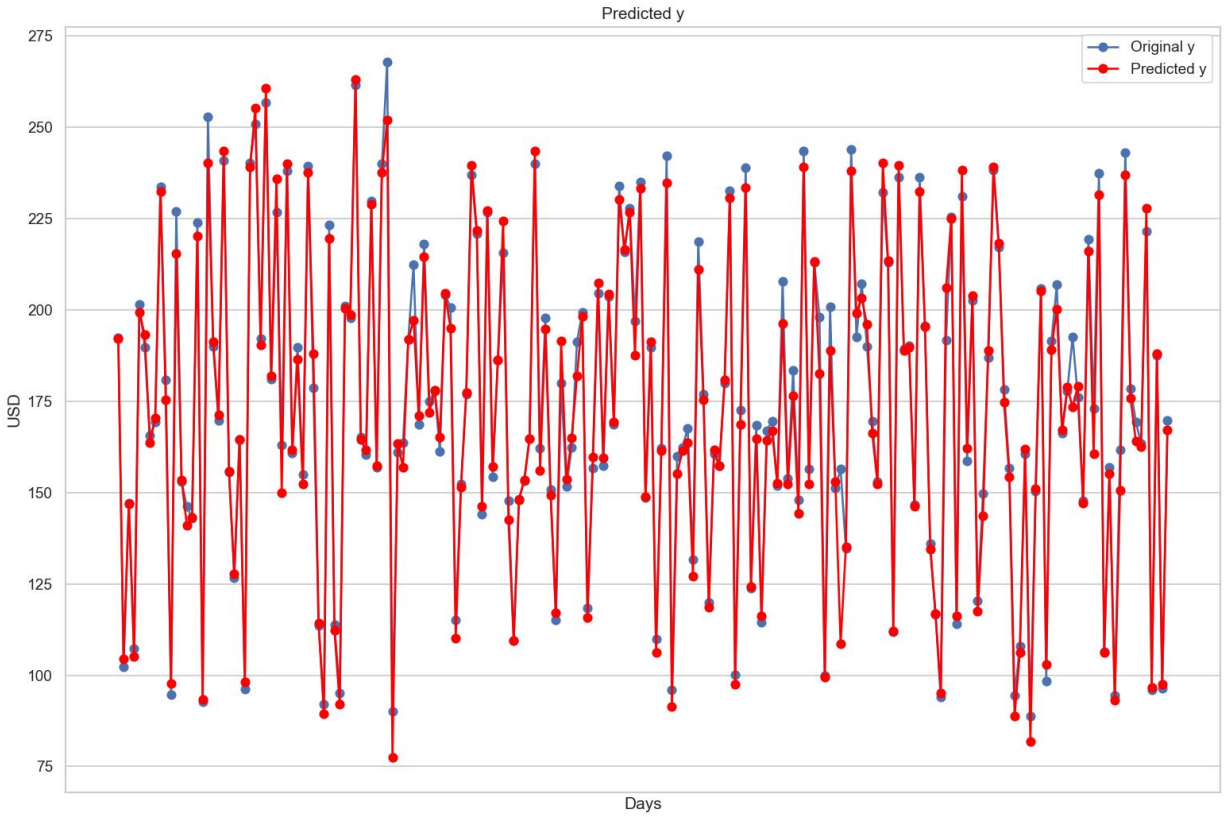


Figure 13: Original stock prices vs Predicted stock prices for Goldman Sachs Group on Test set with best parameters gamma and λ

From the plot we see that prediction for the daily closing price stock is quite close. We can conclude that the model with best parameters provides quite good results for prediction. RMSE and RMSE ratio significantly reduced compared to the previous model. For each case we can clearly see that there is very small difference between true and predicted values of response y.

We identify the 10 cases in the TEST set for which the squared prediction error is the largest the "bad" cases from the following dates:

[2017-03-03, 2011-01-04, 2018-12-28, 2018-12-03, 2010-12-07, 2019-08-27, 2016-11-09, 2015-07-20, 2018-01-31, 2018-12-24]

To visualize 10 "bad" cases we perform Principal Component Analysis and projecting all the TEST cases onto the first 3 PCA.

DARIA KUDOIAR
dariakudoiar@gmail.com

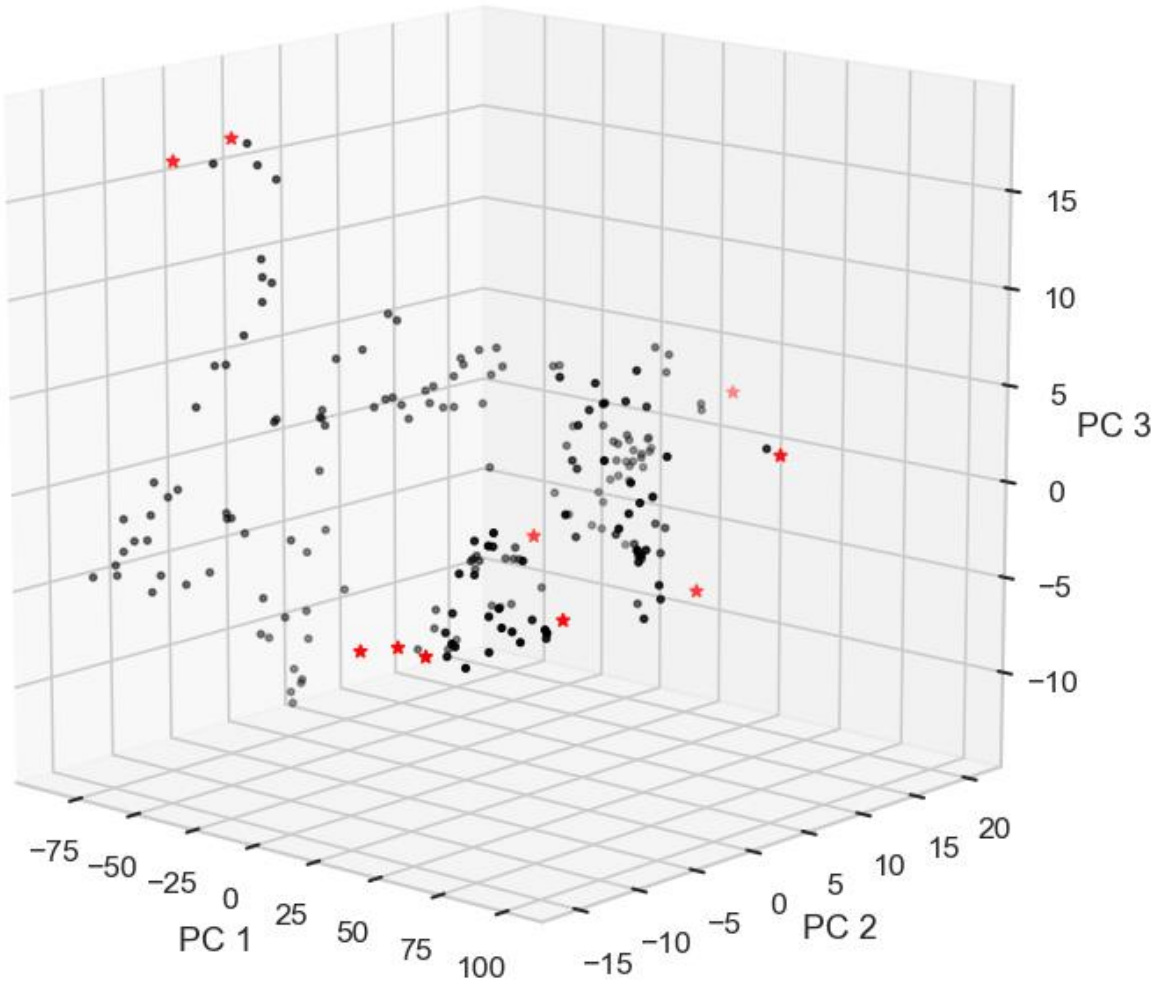


Figure 14: 3D visualization with PCA of cases with the highest prediction error in Test set (red points represent the cases with the highest error)

DARIA KUDOIAR
dariakudoiar@gmail.com

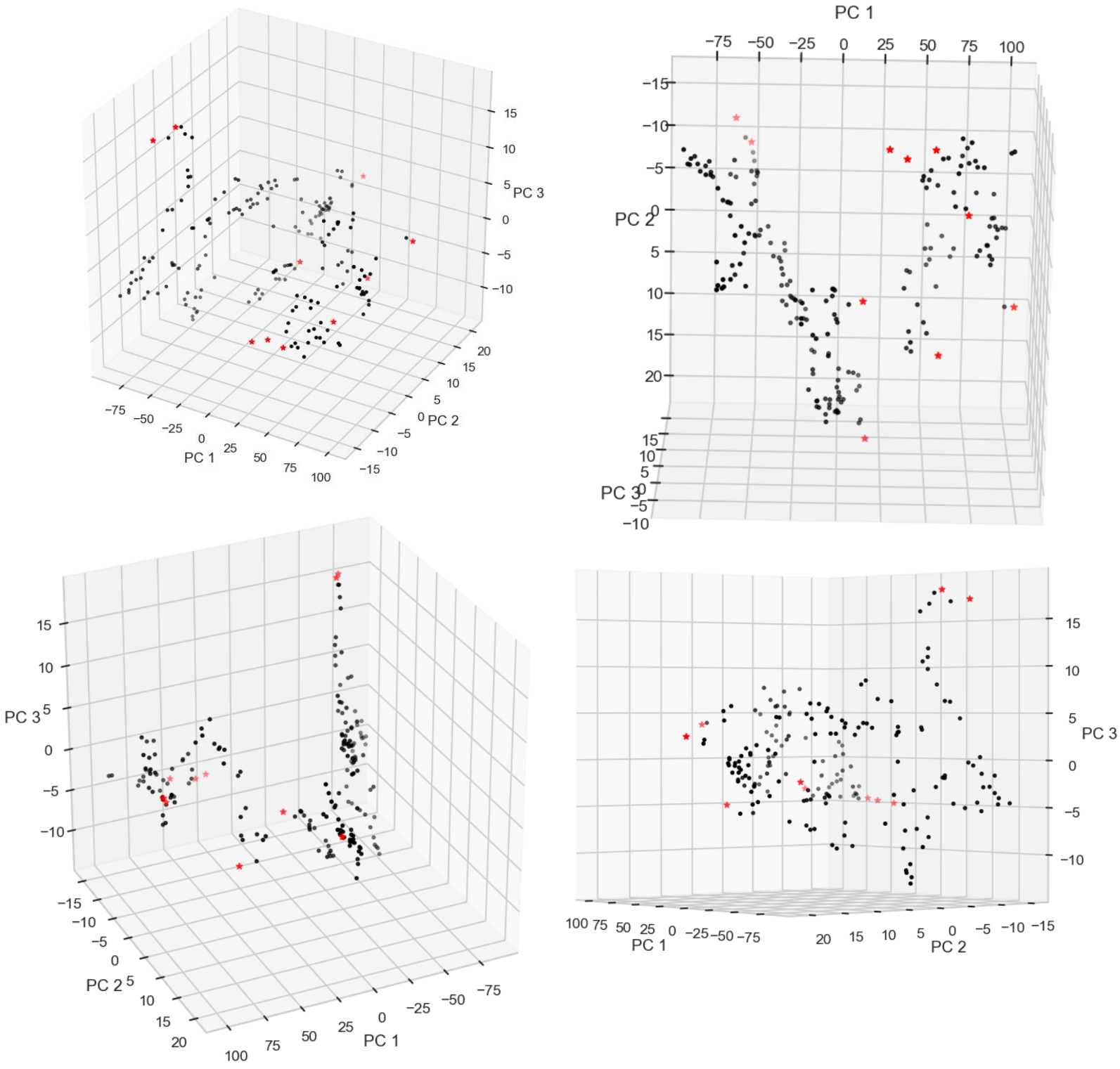


Figure 15: 3D view from different angles on cases with the highest prediction error in Test set (red points represent the cases with the highest error)

DARIA KUDOIAR
dariakudoiar@gmail.com

By look at the 3D PCA plots we can say that ‘bad’ cases somewhat grouped together in a few specific locations. We can clearly see a few outliers somewhere between -50 to -75 for PC2. Since some of the “bad” case mostly confined in a certain area of the new projected space it means they have similarities, but ‘bad’ cases are not far away from some called ‘good’ cases and it seems that we cannot get better separation ‘good’ cases from ‘bad’ cases.

We set closing ‘Dates’ stock prices as index at the beginning of our analysis. We want to see and try to find pattern and compare cases with high and low RMSE.

Indices (dates) for cases with the highest error and lowest errors are:

INDEX of case with high RMSE	*RMSE*	*INDEX of case with low RMSE*	*RMSE*
2010-12-07	420.2	2011-10-26	0.000087
2011-01-04	330.1	2012-07-18	0.02
2015-07-20	674.7	2012-09-25	0.06
2016-11-09	493.4	2013-05-09	0.08
2018-12-03	414.5	2014-08-27	0.01
2018-12-24	2219.7	2016-09-02	0.03
2018-12-28	386.8	2016-06-03	0.003
2018-01-31	1642.9	2017-10-19	0.02
2017-03-03	315.7	2018-03-01	0.0006
2019-08-27	468	2019-01-10	0.01

It seems that 4 out 10 cases with the highest error collected from December 2018/2011. One of the ‘bad’ cases shows extremely high RMSE.

It doesn’t seem that ‘bad’ and ‘good’ cases have clear pattern. However, it seems that more ‘bad’ cases collected from the period when the year ends and begins, December and January.

DARIA KUDOIAR
dariakudoiar@gmail.com

QUESTION 4

To analyze the best predicting formula pred(x) we fix the best choice of parameters and reorder the coefficients |A1|, |A2|, ... , |Am| in decreasing order, to get a list B1 > B2 ... > Bm >0. We plot the decreasing curve B_j vs j.

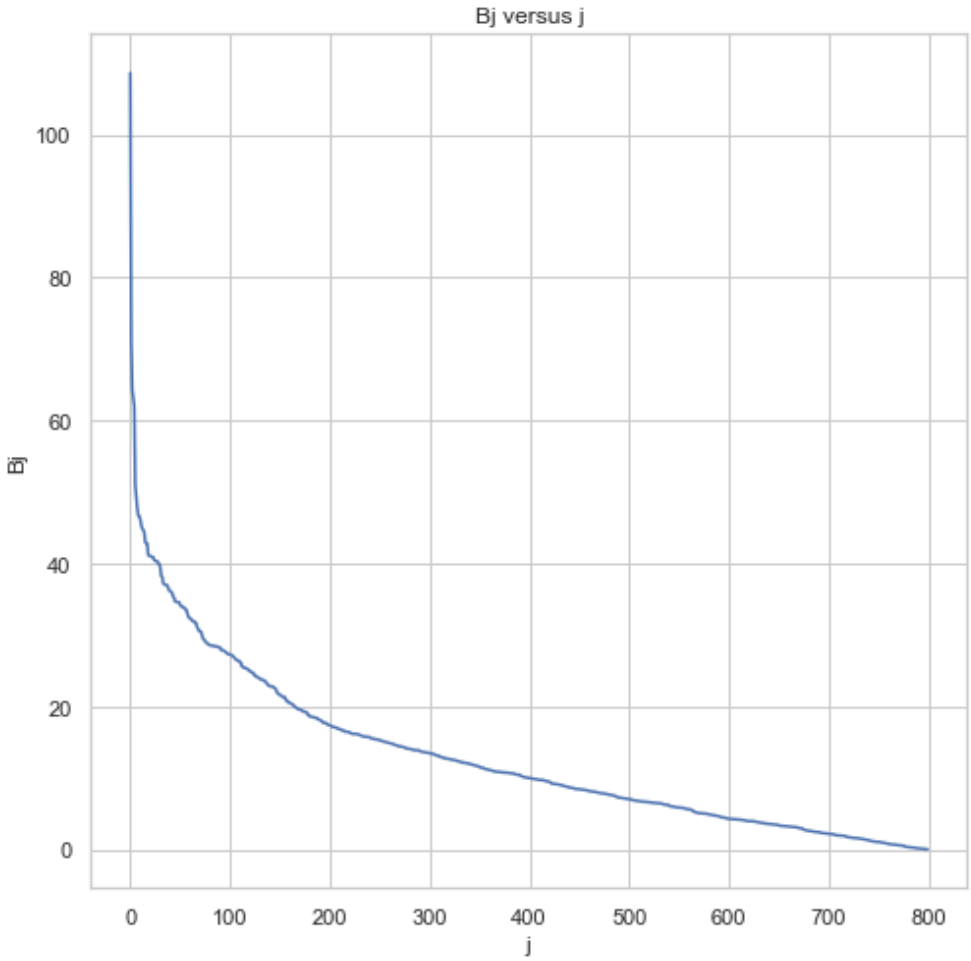


Figure16: Decreasing curve Bj versus j

In addition we compute the ratios bj = (B1 + ... + Bj)/(B1 + ...+Bm) and plot the increasing curve bj versus j.

DARIA KUDOIAR
dariakudoiar@gmail.com

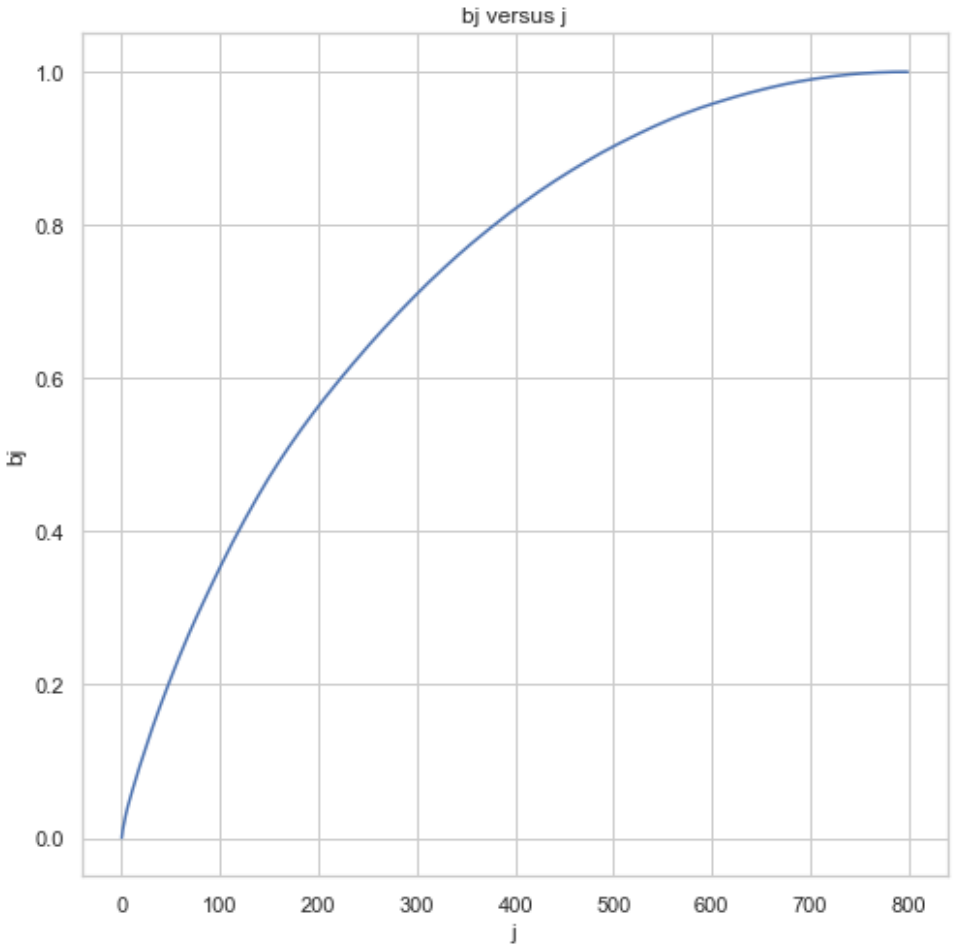


Figure 17: Increasing curve bj versus j

Compute the smaller j = 680 such that $b_j > 99\%$ and the corresponding threshold value THR = $B_j = 0.58$.

We use reduced prediction formula such that for i =1... m, if $|A_i| >$ THR, we fix $AA_i = A_i$ and otherwise set $AA_i = 0$. This yields a reduced formula

$$pred(x) = AA1\ K(x, X(1) + \cdots + AAm\ K(x, X(m)$$

We run the reduced formula on both Train and Test set and evaluate performances.

DARIA KUDOIAR
dariakudoiar@gmail.com

Data	*RMSE*	*RMSE ratio*	*95% Confidence interval for RMSE ratio*
Train set	36.8	0.21	[0.19; 0.24]
Test set	39.6	0.22	[0.16; 0.28]

We see that compared to the original pred(x), reduced formula with a smaller number of coefficients provides higher MSE and RMSE ratio such as 36.8/0.21 and 39.6/0.22, respectively. Scatterplot of Predicted y vs Original y also indicates the big difference between the actual points and predicted ones.

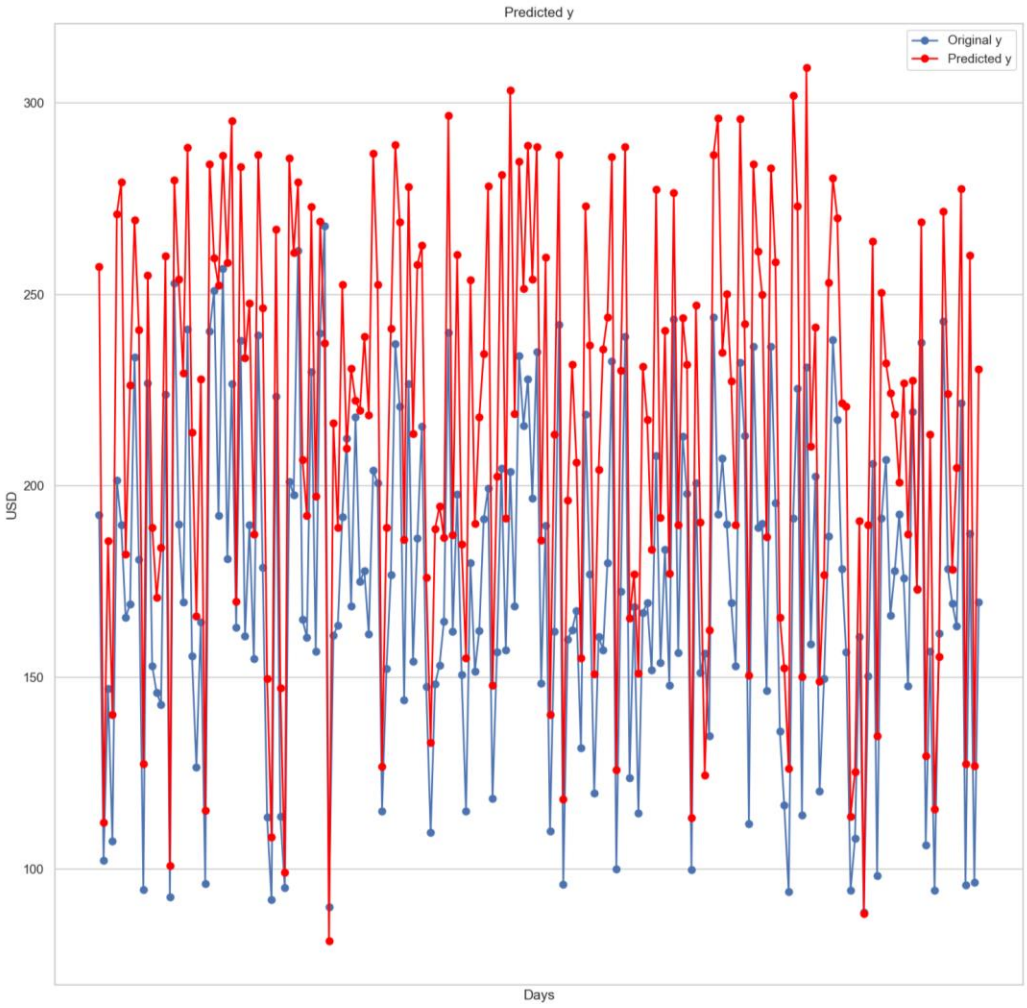


Figure 18: Original stock prices vs Predicted stock prices for Goldman Sachs Group on Test set with reduced prediction formula

DARIA KUDOIAR
dariakudoiar@gmail.com

QUESTION 5

We use pre-existing function in Scikit-Learn Python library to implement Kernel Ridge Regression. We use the best parameters for KRR function gamma = 0.07 and λ = 0.9. Computation time drastically decreased with built in function, from 8 minutes to calculate pred(x) to few seconds, however we get higher RMSE and RMSE ratio for prediction with built-in Python function compared to the original formula and reduced formula.

By implementing KRR precisely 'by hand' with hard code you are able to specifically control each step of implementation. In other hand, built in function where you don't really know what is going inside of it, and there is no way you can determine step by step implementation and the whole process.

We get the following results with KRR function:

Data	*RMSE*	*RMSE ratio*	*95% Confidence interval for RMSE ratio*
Train set	42.23	0.24	[0.21; 0.27]
Test set	70.62	0.27	[0.21; 0.33]

Surprisingly, we see that with KRR() function we have higher RMSE and RMSE ratio compared to the other models implemented before. What is more important, that confidence intervals are not overlapping and there is significant difference between RMSE on Train and Test set and we are confident with our results. Also, we can analyze the Scatterplot of Original stock prices vs Predicted stock prices for Goldman Sachs Group on Test set with built-in KRR() function and it support the conclusion about lower performance with KRR() function compared to the other prediction formulas we`ve used.

DARIA KUDOIAR
dariakudoiar@gmail.com

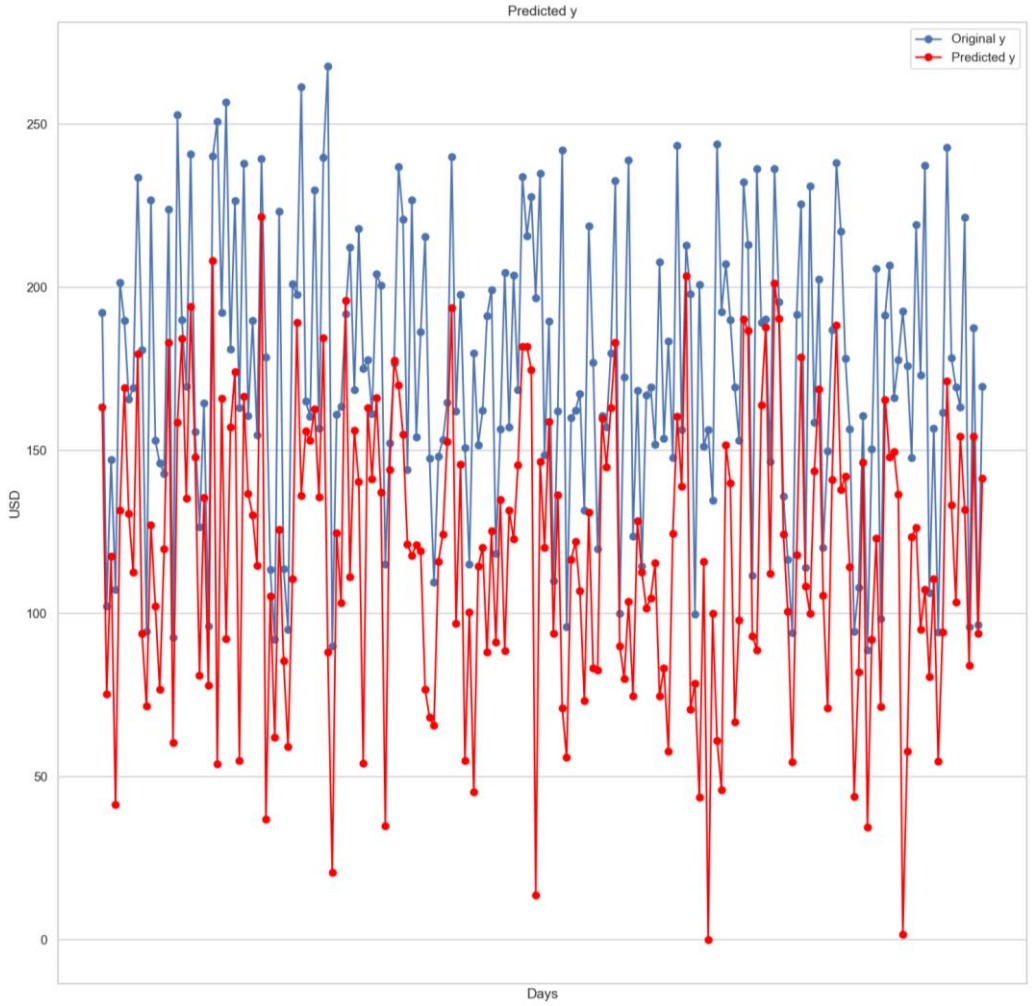


Figure 19: Original stock prices vs Predicted stock prices for Goldman Sachs Group on Test set with built-in KRR function

Scatterplot indicates significantly high difference between some of points for predicted value of y and the true value of y. It means there should be large errors (RMSE). However, we can see that for some specific periods of time, there is not a big difference between predicted and true values of y.

DARIA KUDOIAR
dariakudoiar@gmail.com

CONCLUSION:

The main goal of the project: apply Kernel Ridge Regression (KRR) to predict the value of daily closing stock prices for Goldman Sachs Group Inc whenever a new case X = [X1... Xp] is given. We`ve compared originally defined prediction formula as

$$pred(x) = A1\ K(x, X(1) + \cdots + Am\ K(x, X(m)),$$

reduced formula with smaller number of coefficients:

$$pred(x) = AA1\ K(x, X(1) + \cdots + AAm\ K(x, X(m))$$

and the performance of prediction with pre-existing KernelRidge() function.

We got the following results:

Data	*RMSE*	*RMSE ratio*	*MODEL*
Train set	5.13	0.03	Pred(x) with best parameters
Test set	7.91	0.05	
Train set	36.8	0.21	Reduced Pred(x)
Test set	39.6	0.22	
Train set	42.23	0.24	KernelRidge() function
Test set	70.62	0.27	

We can conclude that Model with originally defined formula provides the best prediction for daily closing stock prices for Goldman Sachs Group Inc. Scatterplots with true values of response versus predicted values supports this conclusion. The best situation we see explicitly with formula Pred(x) with best parameters.

DARIA KUDOIAR
dariakudoiar@gmail.com

We should take into account that our project has the limitation of not considering data quality assurance methods. Some results presented here suffer from poor data inputs and future studies should consider data treatment before usage. The present research approach equates missing minute prices to the previous values, not considering contiguous missing minutes or interpolating values. Some of the selected stocks illustrate the influence over the results of long streaks of missing data as well as outliers. In addition, markets (stocks) generate data that form (statistically) non-stationary, time-series of numbers over any period of 'time window' that one may want to examine. Prediction, which is highly 'precise', is essentially impossible, but to a greater or lesser degree, less precise, but more probabilistic prediction could be applicable to market time-series data.