PREDICTION OF STOCK MARKET PRICES OF HEALTH SECTOR COMPANIES

- KISHORE TUMARADA
- GRADUATE STUDENT IN STATISTICS AND DATA SCIENCE

PROBLEM STATEMENT

Goal 1 : Predict the increasing or decreasing trend of closing stock price of a company on a day.

Goal 2 : Predict the stock price of a company.

GOAL 1 : PREDICTING TREND OF STOCK PRICE

- Trend here is considered as moving average of the closing stock price for previous 5 days.

DATASET DESCRIPTION

- Ten companies are selected from Health sector with following conditions:
  - Target company Gilead Sciences' closing price has high correlation with remaining 9 companies' prices.
  - But remaining 9 companies have low correlation among themselves.
- Based on above conditions, closing stock prices from 01/01/2015 to 12/31/2019 for 10 companies are extracted from yahoo finance.

S NO	COMPANY NAME	TICKER SYMBOL
1	Gilead Sciences	GILD
2	Edwards Lifesciences	EW
3	Boston Scientific	BSX
4	DaVita Inc.	DVA
5	Alexion Pharmaceuticals	ALXN
6	AmerisourceBergen Corp	ABC
7	Regeneron Pharmaceuticals	REGN
8	Waters Corporation	WAT
9	Intuitive Surgical Inc.	ISRG
10	Danaher Corp.	DHR

Table 1. companies in dataset

DATASET SAMPLE

- Top 5 rows of original dataset :

	Date	GILD	EW	BSX	DVA	ALXN	ABC	REGN	WAT	ISRG	DHR
0	2015-01-02	94.910004	63.860001	13.22	75.830002	186.600006	90.459999	410.160004	113.879997	175.190002	64.988625
1	2015-01-05	96.790001	63.889999	13.81	74.699997	182.169998	89.690002	412.470001	113.019997	171.456665	64.344200
2	2015-01-06	97.650002	63.509998	13.70	73.620003	177.949997	90.180000	396.890015	112.529999	173.263336	63.904472
3	2015-01-07	99.480003	65.000000	14.03	74.290001	187.929993	91.980003	407.720001	115.930000	174.213333	64.291130
4	2015-01-08	102.300003	66.574997	14.59	75.769997	183.800003	92.190002	403.250000	118.089996	177.666672	65.284309

Table 2. Head of total dataset

DATASET DESCRIPTION

- Features are created by calculating moving averages calculated for 10 companies, including the target company, for past 10 days. Hence a total of 100 features are generated to predict the trend.
- The trend of target company is considered as classes with class 1(if increasing) and class (-1)(if decreasing).
- This dataset is divided into train and test data in ratio of 80% to 20%.
- The data along with moving average and trend is shown in Fig 1.

Fig 1. Time series plot of GILD stock prices(blue), 5 day moving average(orange) and trend line(green)

SVM FOR PREDICTING TREND

- Support Vector Machine with polynomial kernel has been trained on this dataset.
- For polynomial SVM, kernel is $K(x, y) = (coef0 + < x, y >)^5$
- Performance of best polynomial SVM (with cost C = 0.001, coef0 = 15, gamma =1) on test data can be seen by confusion matrix and ROC plots for CL 1 and CL -1.

	pred_CL1	pred_CL(-1)
true_CL1	69%	31%
true_CL(-1)	33%	67%

Table 3. confusion matrix for test data

RESULTS

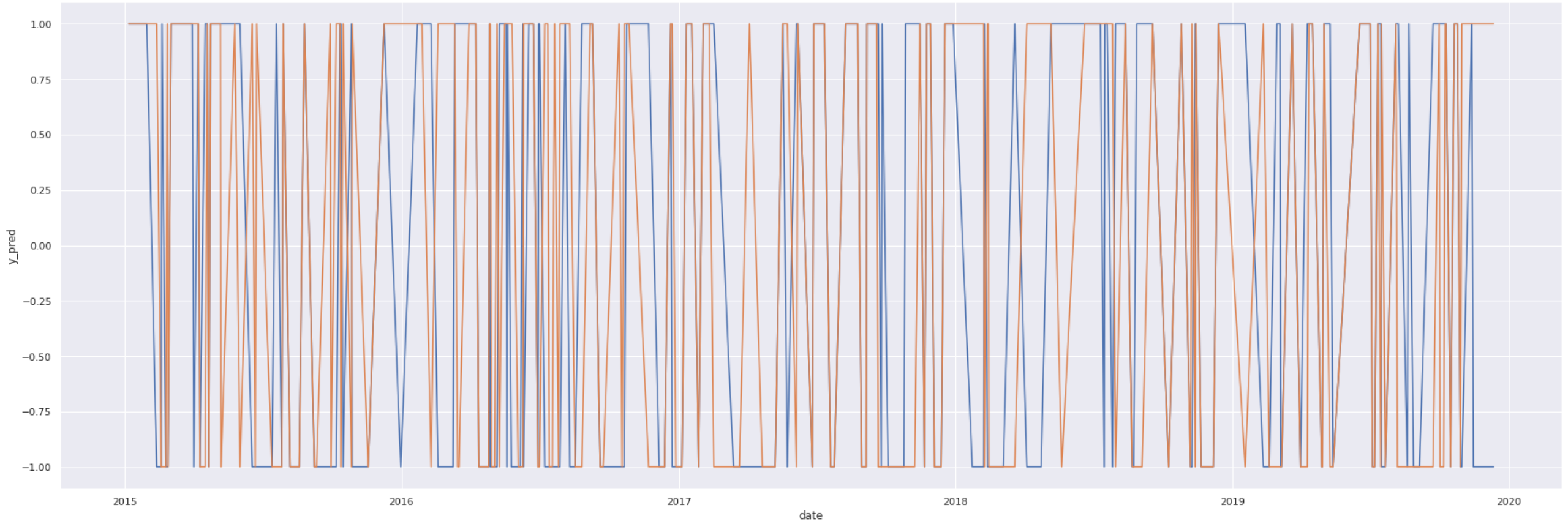

Fig 2. Time series plot of trend lines – true trend(blue) and predicted trend(orange)

PERFORMANCE

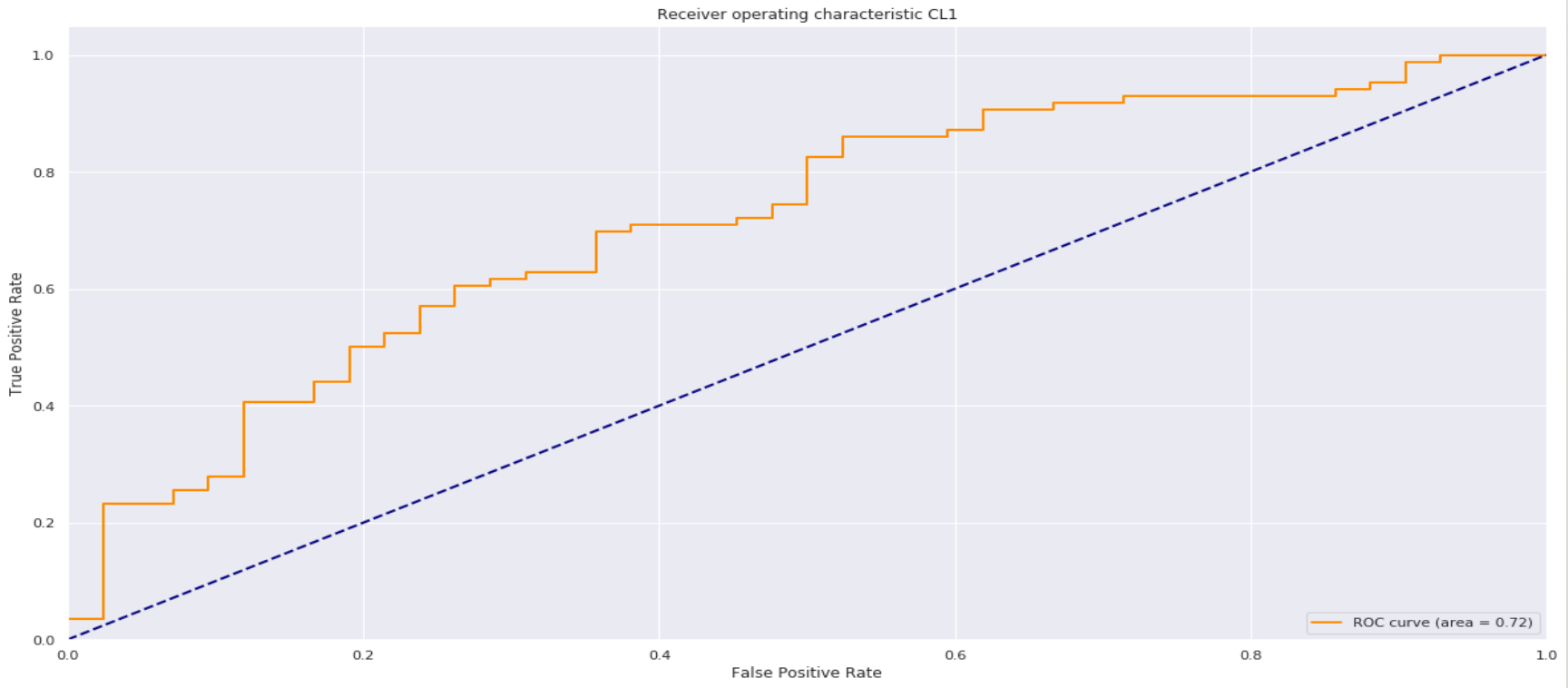

Fig 3. ROC plot for Class 1 with AUC =0.72

PERFORMANCE

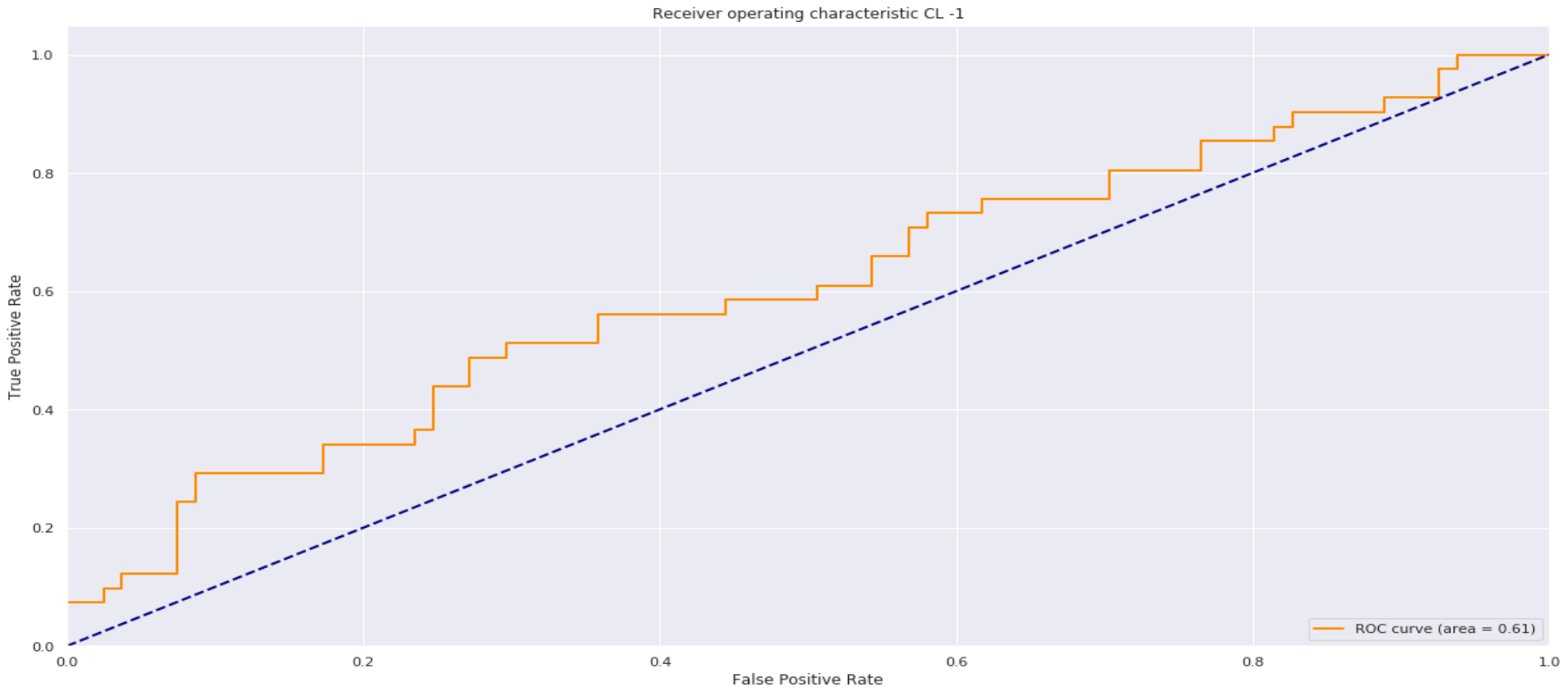

Fig 4. ROC plot for Class -1 with AUC =0.61

TECHNICAL COMMENT

- The support vector ratio is 0.51, which is quite high. It implies that the model is taking support of 51% of train data to predict the remaining 49% cases of train data.
- ROC plots in Fig 3 and 4 also tells the same story as confusion matrix with higher prediction rate for class 1 than class (-1).

GOAL 2 : PREDICTING STOCK PRICE

- Predict the closing stock price of one company on a day based on closing stock prices of previous 10 days prices of 10 companies including the target company.

DATASET DESCRIPTION

- Dataset is same as previous problem.
- A new dataset is generated with 100 features/explanatory variables – past 10 day closing stock prices of 10 different companies including target company(Gilead Sciences).
- The target/dependent value is closing stock price of target company on next day.
- This dataset is divided into train and test data in ratio of 80% to 20%.

S NO	COMPANY NAME	TICKER SYMBOL
1	Gilead Sciences	GILD
2	Edwards Lifesciences	EW
3	Boston Scientific	BSX
4	DaVita Inc.	DVA
5	Alexion Pharmaceuticals	ALXN
6	AmerisourceBergen Corp	ABC
7	Regeneron Pharmaceuticals	REGN
8	Waters Corporation	WAT
9	Intuitive Surgical Inc.	ISRG
10	Danaher Corp.	DHR

Table 4. companies in dataset

KERNEL RIDGE REGRESSION TO PREDICT PRICE

- A kernel ridge regression model, with radial kernel $\mathbf{K}(\mathbf{x}, \mathbf{y}) = \mathbf{e}^{-\boldsymbol{\gamma} * (\|\mathbf{x}-\mathbf{y}\|)^2}$, is trained on this dataset. The parameters tuned are cost λ and gamma γ.
- Methodology followed for tuning of parameters is :
  - After selecting a baseline values, in each step, only one of the parameter is tuned with 2 combinations –
    - When $\lambda = \lambda_0, \gamma = [\frac{\gamma_0}{2}, 2 * \gamma_0]$
    - Similarly when , $\gamma = \gamma_0, \lambda = [\frac{\lambda_0}{2}, 2 * \lambda_0]$
  - In each step the combination of parameters, for which rmse for test and train are low and have less difference between the errors , are chosen.
- Best parameters after parameter tuning are : λ = 0.0156, γ = 0.0001 .

RESULTS

Fig 5 Time series plot of true y and predicted y values

WORST CASE ANALYSIS

- Performance is measured by Root mean square error, which is
  - $RMSE = \sqrt{\frac{\sum_1^k (y - \hat{y})^2}{k}}$, where y is true target variable and $\hat{y}$ is predicted target.
  - $ratio = \frac{RMSE}{avy}$, where $avy = \frac{\sum_1^k |y_i|}{k}$, i.e., mean of true values of target variables. This normalizes the rmse value with average of true y
- Performance measures are rmse = 4.932, ratio rmse/avy : 0.0611.
- It implies that prediction error percentage is 6%.

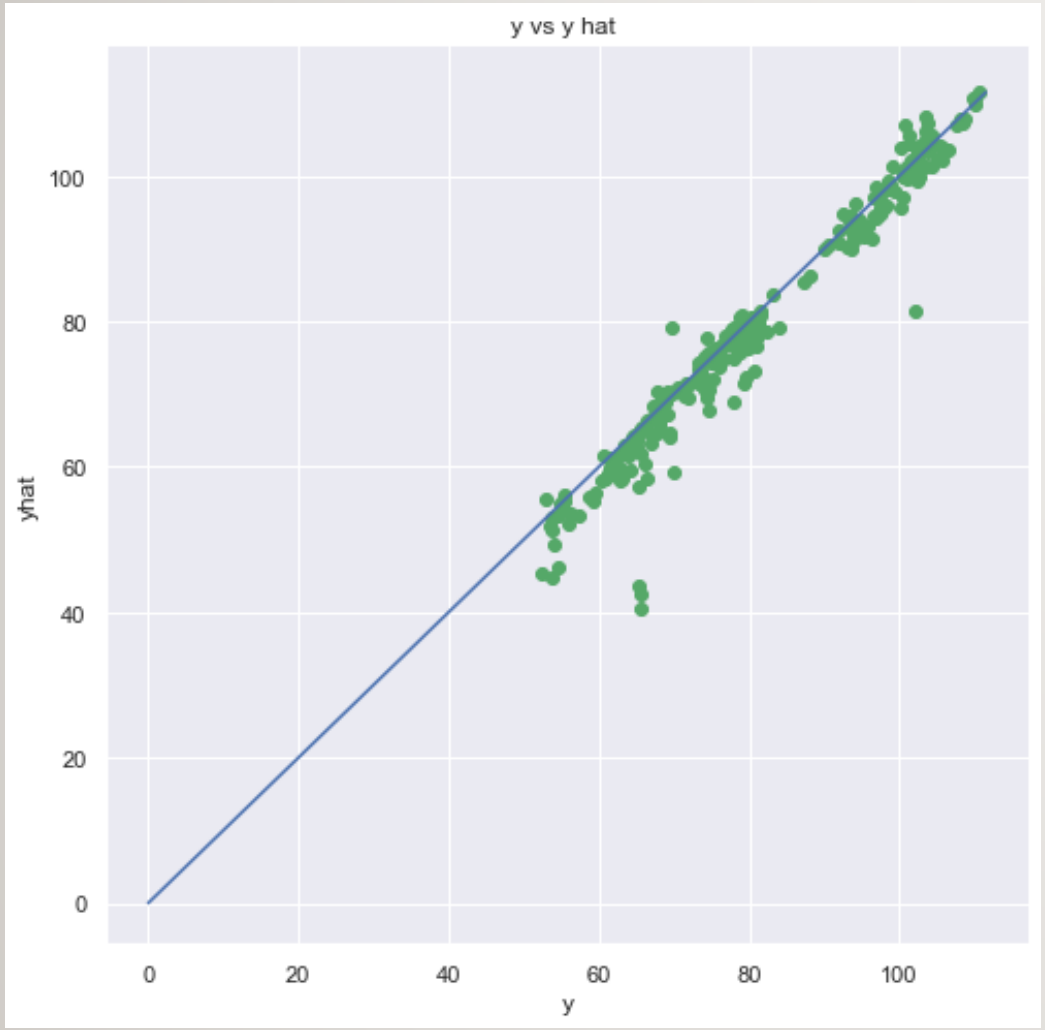

Fig 6. scatter plot of true y and predicted y for test data

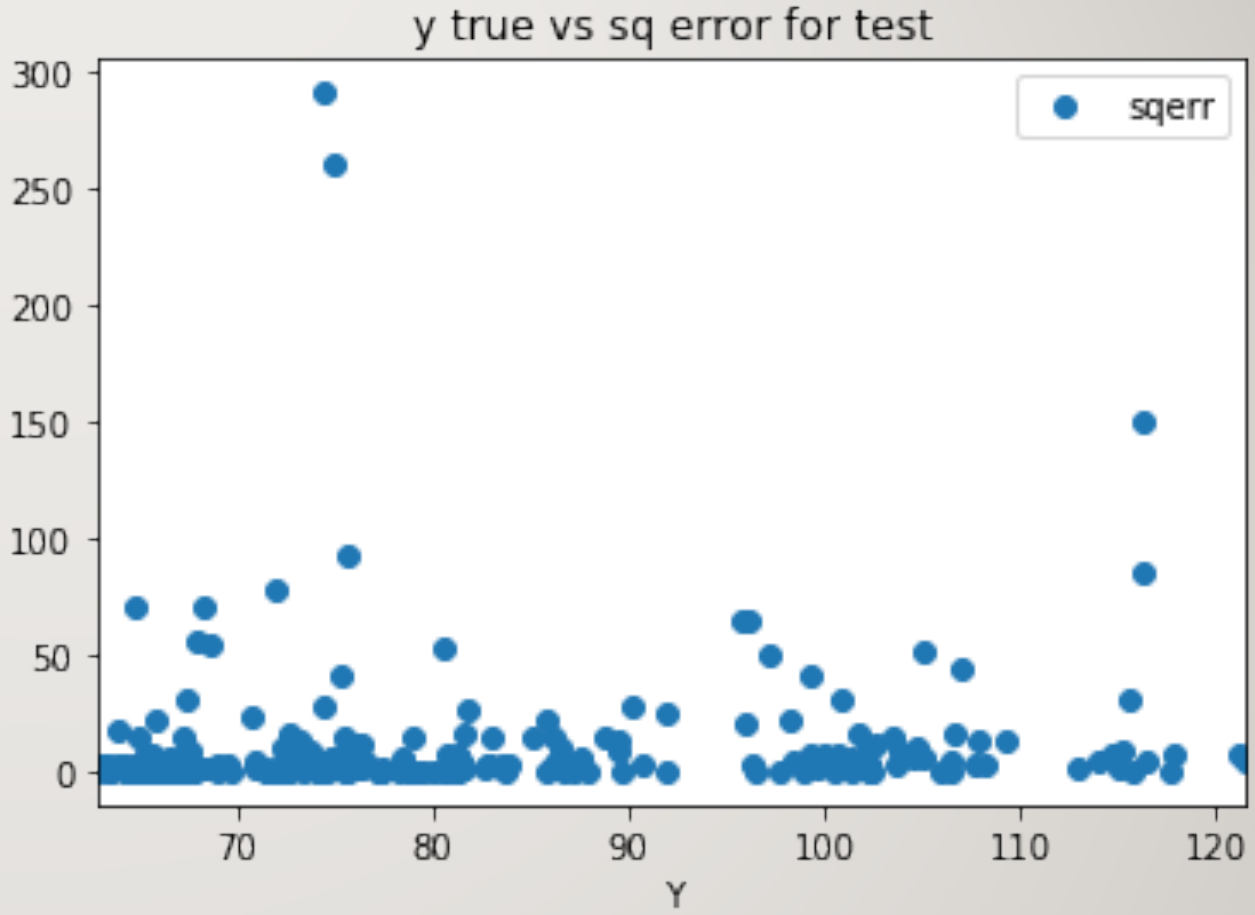

Fig 7. Scatter plot of true y and square error for test data

WORST CASE ANALYSIS

- The Figure 7 indicates that error in prediction is higher for y true values in between 60 and 75$; between 95$ and 105$ stock prices.
- In Fig 8, top 10 best cases(dark blue) and 10 worst cases(red to light blue) in terms of squared error are shown by projecting first three Principle Components after PCA. It indicates that, worst cases are concentrated in PC1>200 region.

WORST CASE ANALYSIS

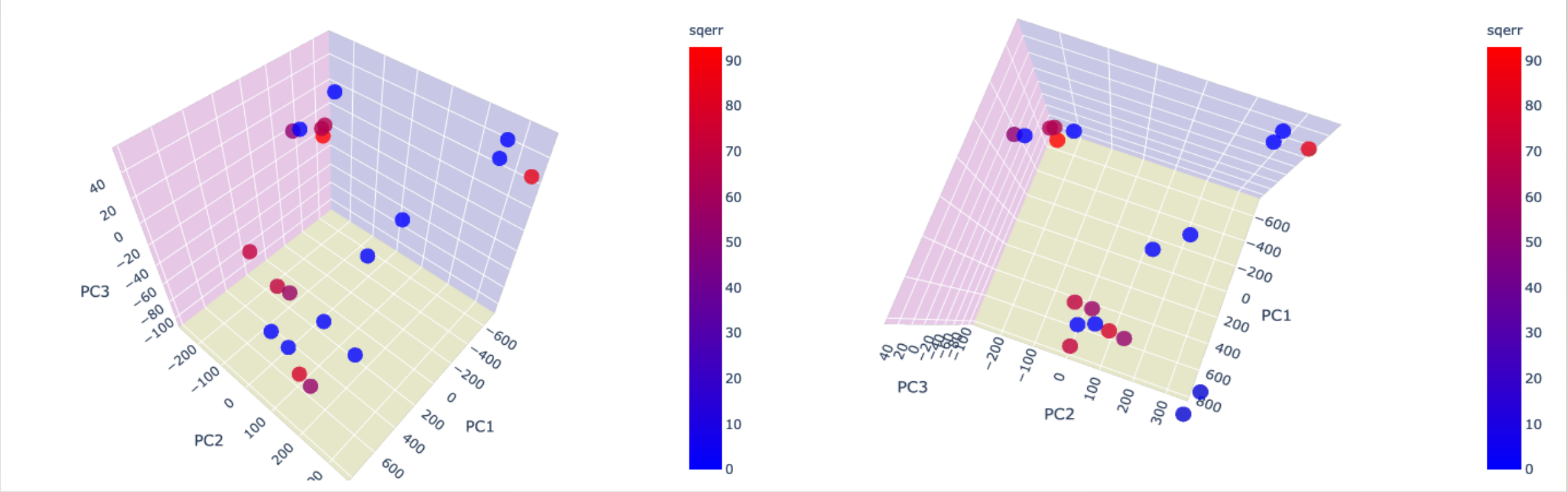

Fig 8. PCA analysis of 10 best cases(blue) vs 10 worst cases(red) in terms of square error

CONCLUSION

- The KRR model is able to predict closing stock price for next day for the target company(GILD) with an error percentage of 6%.