SUPERVISED LEARNING: REGRESSION

Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation.

Main objective with this work is to model for prediction purposes.

Benefits:

- Main aim for this project from a business perspective is to increase earnings for the company.
- Since the company is wants to find optimum price point to maximize their earnings, it tries to predict based on past data if a product's based on product, demand and purchase details.

Brief description of the data set you chose and a summary of its attributes.

The data I took is about pricing a new product.

So, the data is made up of 284780 rows (each of a distinct applicant) and 7 features for prediction and one target variable of which training data has 227820 entries, and test has rest. So we have 284780 instances of products sold across invoices and stockings and what price they were sold at. The features/ variables and their variable types are:

- Date: InvoiceDate
- Numerical: Quantity
- Categorical: InvoiceNo, StockCode, Description, CustomerID, Country

I am trying to model the data to predict whether what a product's price will be based on these features.

Brief summary of data exploration and actions taken for data cleaning and feature engineering.

1. Handling date variables
    a. Since we cannot handle date variables directly, we can get the information out of it by creating new columns.
    b. To see if date or month or year of Invoice Date has any effect in determining target, creating separate columns each for date, month and year instead of datetime format directly
2. Handling non-numerical variables:
    a. Frequency encoding was performed on Country and StockCode
    b. Invoice number was to give info on what products were purchased together and had no impact on price determination. Hence it was removed.
    c. Target encoding was performed on Description, CustomerID
    d. Invoice had only 2 years 2010,2011. Hence it was binary encoded.
3. Handling missing values
    a. There was no missing data.
4. Handling outliers

a. Flooring and capping were performed on Quantity
b. Target was restricted to 95 percentile

5. Variable Transformation
    a. Boxcox transformation was applied on Quantity
    b. A new column was created to mark Quantity quantile due to its vast range.
    c. Target was also extremely vast in its range, so it was quantile cut, mean calculated for each bin and then log transformed.

Summary of training at least three linear regression models which should be variations that cover using a simple linear regression as a baseline, adding polynomial effects, and using a regularization regression.

The modelling methods which I too into consideration were: Simple Linear regression, Polynomial Regression, Ridge, Lasso, Bayesian Ridge, Polynomial Ridge, Support Vector Machine. Note that all of these we modelled and predicted on same train & test data. Hyperparameters and final results were obtained through grid search and 10 fold cross validation.

The exact hyper-parameters used for each method were:

1. ***LinearRegression****(n_jobs=-1)*
2. ***LinearRegression****(n_jobs=-1).fit(**PolynomialFeatures**(degree = 3)*
3. ***Lasso****(alpha = 0.01, max_iter=500,selection='random')*
4. ***Ridge****(alpha=0.01, max_iter=500,solver='lsqr')*
5. ***Ridge****(alpha = 0.1, max_iter=500,solver='lsqr').fit(**PolynomialFeatures**(degree = 3))*
6. ***BayesianRidge****(alpha_1=1.0, alpha_2=1e-6, lambda_1=1e-6,lambda_2=1.0,n_iter=500)*
7. ***SVR(****C=1,gamma=0.1,tol=0.001, max_iter=500)*

The results I got were:

----------Training Data----------

	explained_variance_score	mean_squared_error	mean_absolute_error	max_error	r2_score
Linear regression	0.844631	0.012719	0.083342	1.140415	0.844631
Poynomial regression	0.952611	0.003879	0.039438	1.016423	0.952611
Lasso	0.818051	0.014895	0.091205	1.010555	0.818051
Ridge	0.844375	0.012740	0.083545	1.139075	0.844375
Poynomial Ridge	0.950457	0.004056	0.040541	1.010776	0.950457
Bayesian Ridge	0.844631	0.012719	0.083343	1.140406	0.844631
SVM	0.731112	0.171393	0.386827	1.119519	-1.093650

----------Test Data----------

	explained_variance_score	mean_squared_error	mean_absolute_error	max_error	r2_score
Linear regression	0.846484	0.012568	0.083066	1.109285	0.846477
Poynomial regression	0.953512	0.003806	0.039002	1.071636	0.953511
Lasso	0.820052	0.014732	0.090740	1.006712	0.820045
Ridge	0.846188	0.012592	0.083255	1.109132	0.846181
Poynomial Ridge	0.952313	0.003904	0.039977	0.939067	0.952312
Bayesian Ridge	0.846484	0.012568	0.083066	1.109286	0.846476
SVM	0.733267	0.170194	0.385430	1.014502	-1.078946

A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.

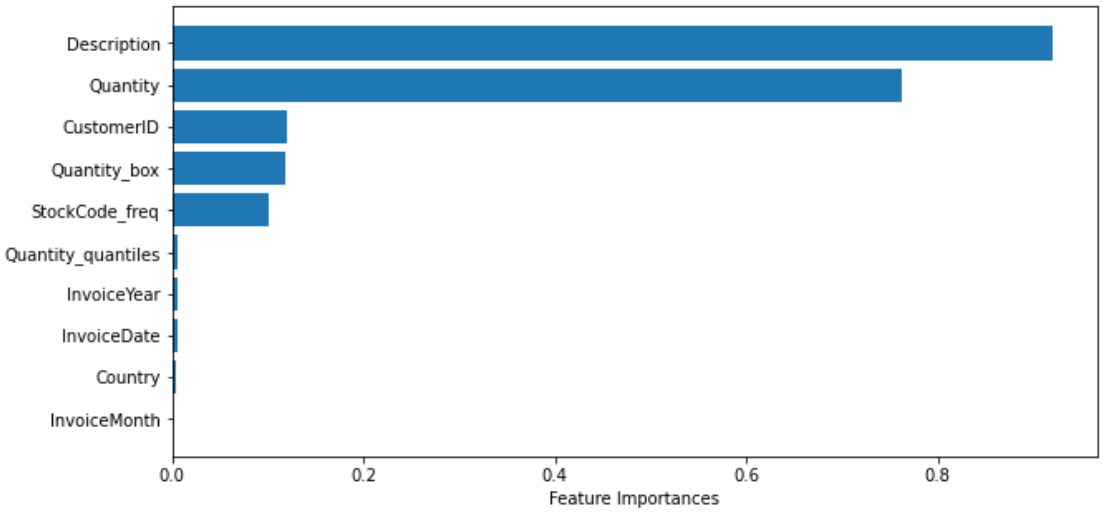
Based on summary statistics of results, we can infer the following:

1. For all the methods, with respect to all comparison metrics, train and test data perform similarly.
2. SVM is worst giving negative R^2 score for linear modelling.
3. Polynomial feature used models perform better than the rest, though they take a lot of computational time.
4. In fitting direct data with no polynomial features, Ridge and Bayesian Ridge are the best, with Bayesian Ridge giving slightly better R^2 score.

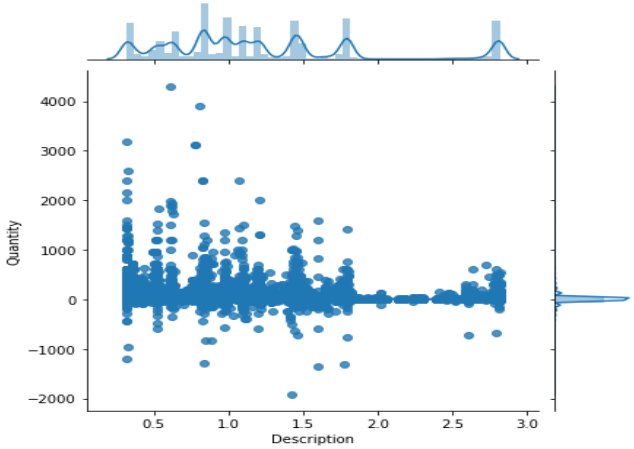
From a perspective of explainability and ease of use, and computational time, Bayesian Ridge is better and with accuracy and model fit Polynomial Ridge is better. I would recommend these two models.

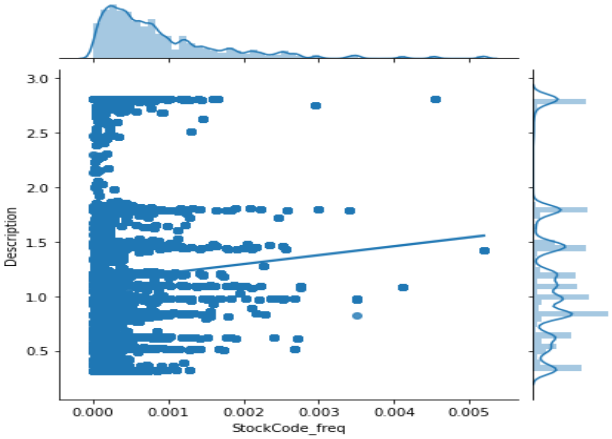
Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your linear regression model.

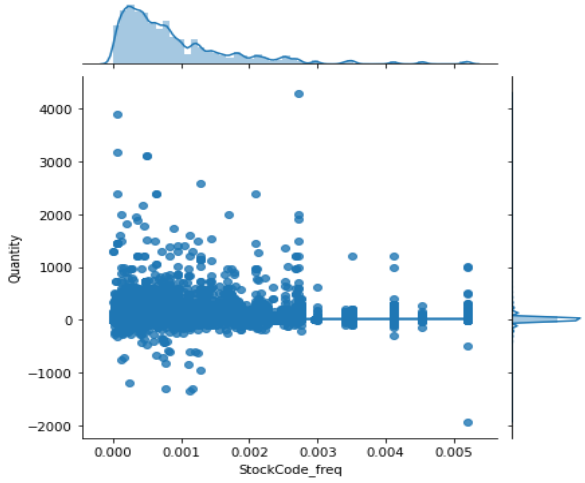
Feature importances based on the final tree is plotted here:

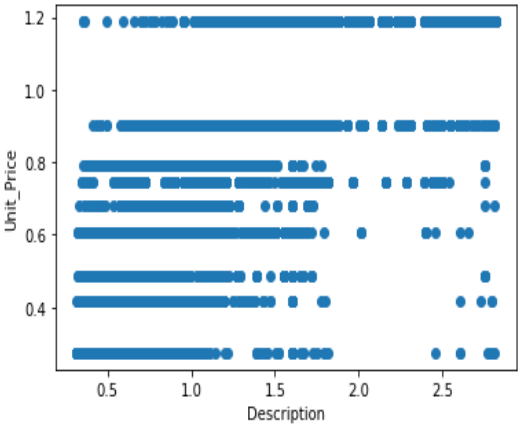


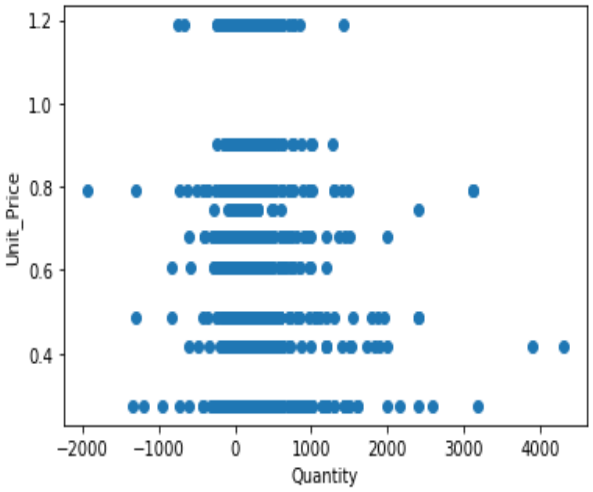
The bivariate plots and joint plots for top few important variable are:

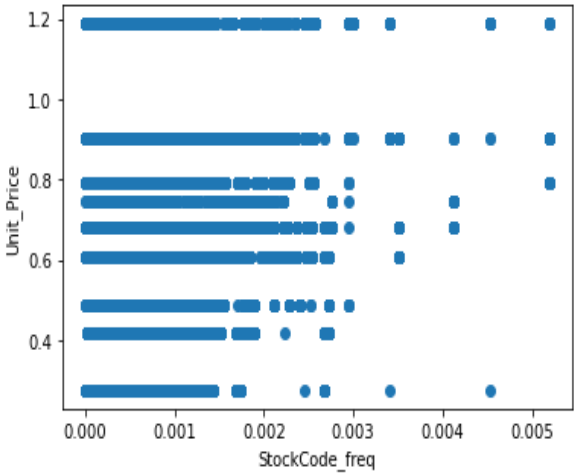












As we can infer from above plot, Description and Quantity are the top 2 primary drivers in determining price of the product with Customer ID and Stock Code being the next 2. Country is one of the least important drivers in determining unit price.

For Polynomial ridge, variables involving the following 5 features were the primary drivers predicting price (in decreasing importance):

1. Description
2. InvoiceMonth
3. Quantity
4. InvoiceYear
5. Quantity_quantiles

Suggestions for next steps in analysing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation or a better prediction.

Next step would be to do feature selection, include dimensionality reduction in case of Polynomial ridge and test for multi-collinearity using VIF.

Additional data regarding cost to manufacture and number of competitors and competitors' profit margin would give better insights into the overall prediction of a product's price.