**IE7275 Final Project Report:Impact of Covid-19 on beer sals**

**-- Team 22**

### **Problem Setting**

Digital technologies, artificial intelligence and machine learning are accelerating in all aspects of our work and life, and are having a significant impact on business management and operations. They are having a significant impact on business management and operations. Many companies are beginning to apply data science in depth to inspire transformative innovation and digital transformation. Advanced technologies can help us extract many features and trends from data that cannot be noticed manually, and apply advanced algorithmic models that can be used to improve the company. The application of advanced algorithmic models can be used to improve business statistics and forecasting, optimize business decisions, and enhance operational efficiency.

### **Problem Definition**

Demand forecasting is the beginning of the entire sales and logistics supply chain planning and is one of the most critical planning operations for any consumer goods company. Proper demand forecasting can effectively guide sales planning and supply chain production planning, package procurement planning and production planning, which is important for business operation efficiency. It is important for the improvement of business efficiency. In the past, demand forecasting has relied more on manual extrapolation and simple traditional time series models. Recently, digitalization transformation has taken root in every industry, and it is imperative to apply advanced algorithmic models to improve the performance of a company's statistical forecasts.

With lots of types of wine, the company separates its products as many brands to aim different customers and areas. Vast information can lead to incorrect sales estimates between manufacturers and dealers. At this time, the processing and prediction of data is very important.With lots of types of wine, the company separates its products as many brands to aim different customers and areas. Vast information can lead to incorrect sales estimates between manufacturers and dealers. At this time, the processing and prediction of data is very important.

### **Data Source**

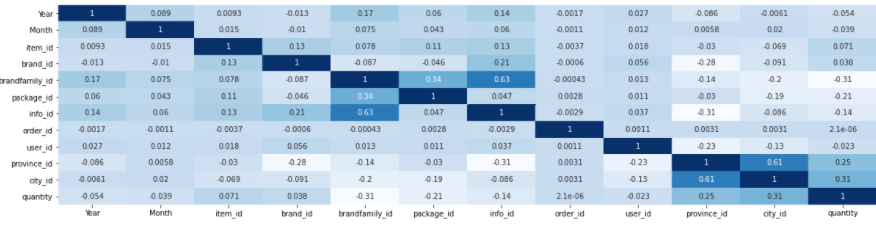
The sales data will be collected from Budweiser and the data will be desensitized.

### **Data Description**

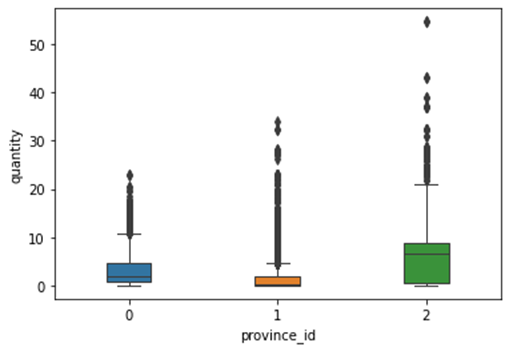
There will be two datasets. The first one will be “item”, it consists of 118 rows and 5 columns. Its column names include columns like item, brand, brand series. The second one will be sales data named “train”, it has 145195 rows and 7 columns. Column names includes columns like date, item, order number, city, and state.

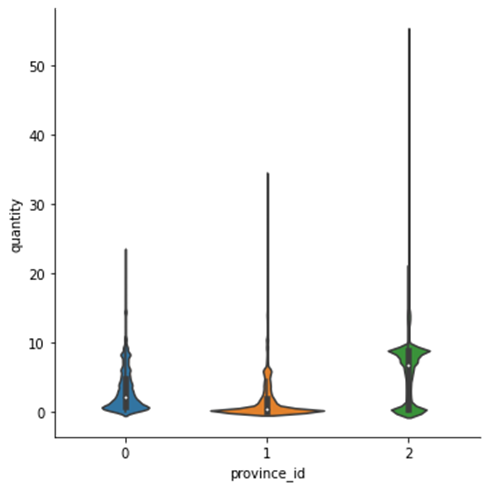
### **Data Exploration**

The first thing to do for EDA is using seaborn to draw the heatmap.

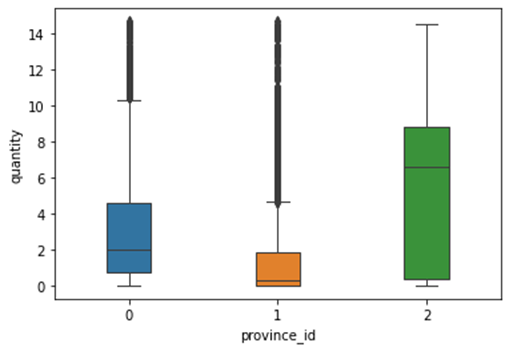
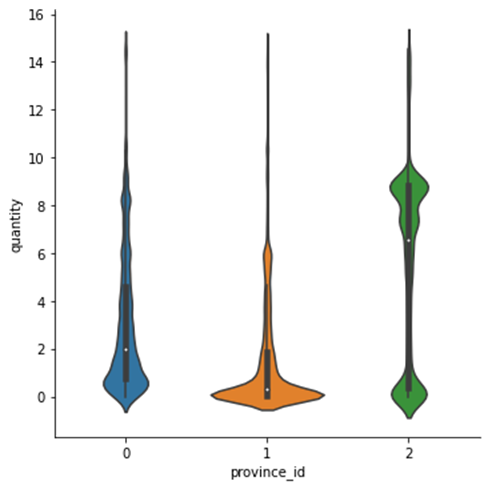


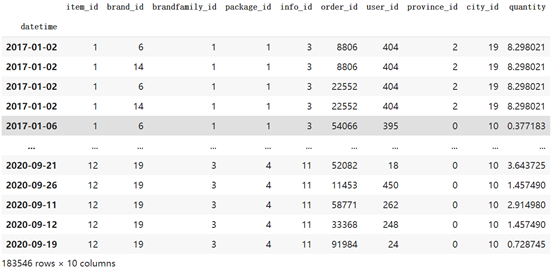
It shows that predictors are almost independent with each other. Most Correlation values are very small and there are correlations greater than 0.7. It is not necessary to do dimension reduction at this stage.

Also, by seaborn, the second step is to draw the boxplot and violin plot. 



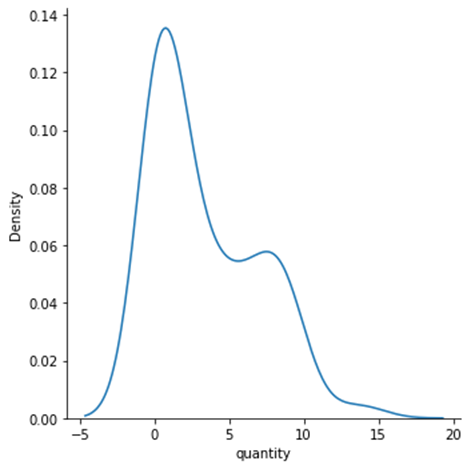
These two results look very weird and there are a lot of outliers, they need to be removed as much as possible.





Set Q1 and Q3 equals to quantile(0.25) and quantile(0.75), apply IQR rules and drop from the index, the rows left are 183546. Although the boxplot and violin plot does not look perfect, they are much better and it’s necessary to keep some outliers for reality.

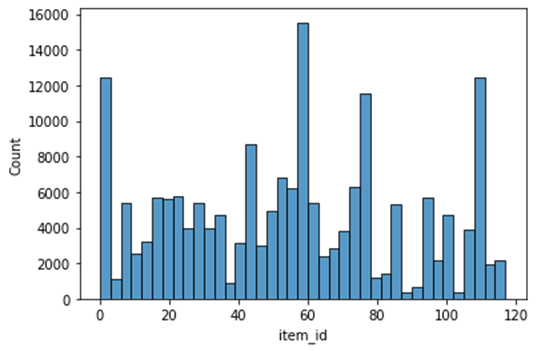
***Data visualization***



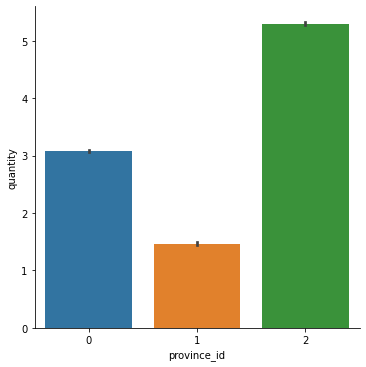
Keep using seaborn package in this part, the first graph is density plot by displot.

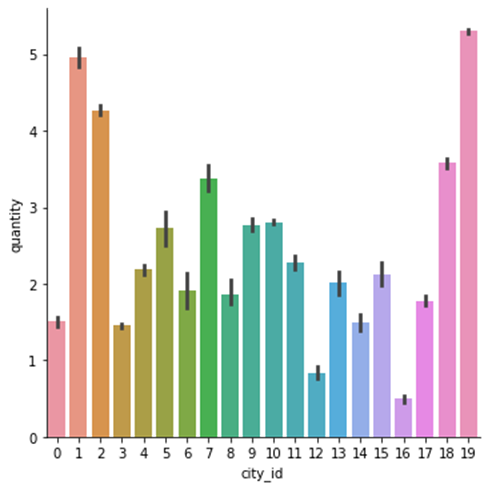
From the density plot of quantity, it shows that most quantities are small and near 0, only a small part are large.

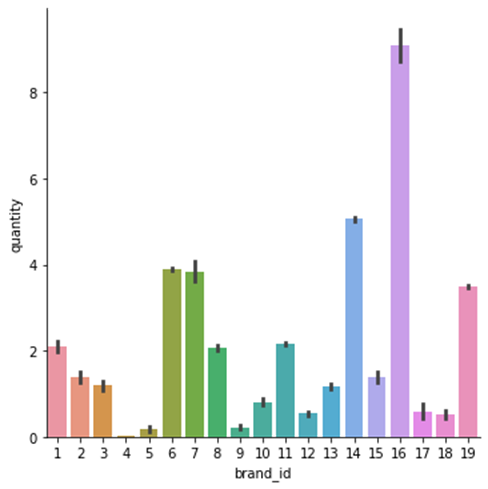
And then histograms by histplot



The histograms are through id and there is a huge difference between each item.

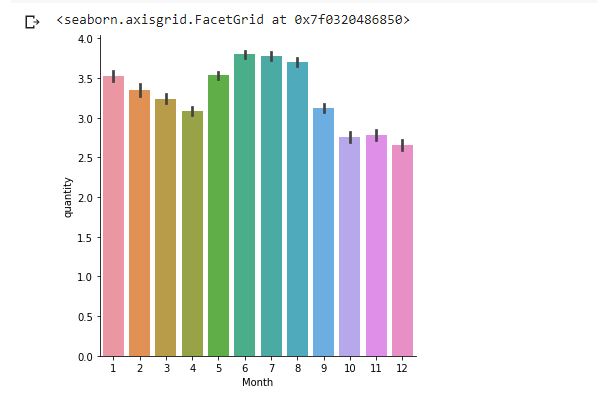
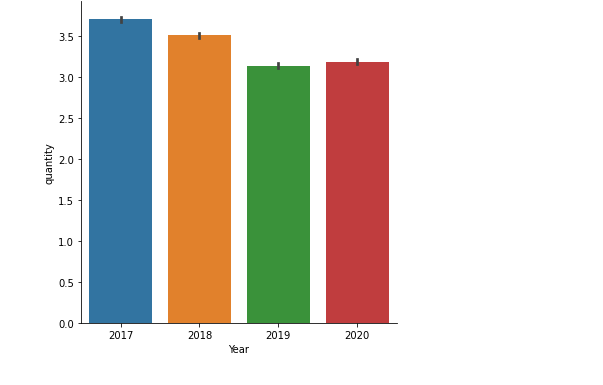


By bar chart it can be found that province 2 has most quantities, followed by province 0. Province 1 has the least quantity.



And the sales performance difference among cities and brands are also huge, which laid the base that the further analysis and prediction should focus on items and states.

Then we draw bar graph with different years and months

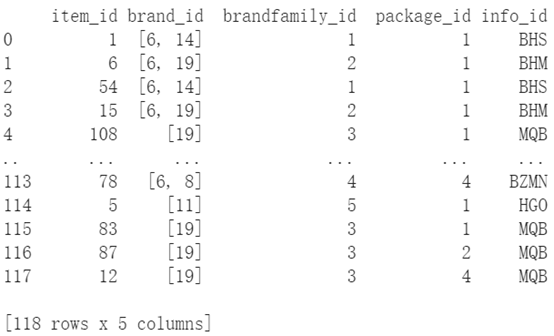


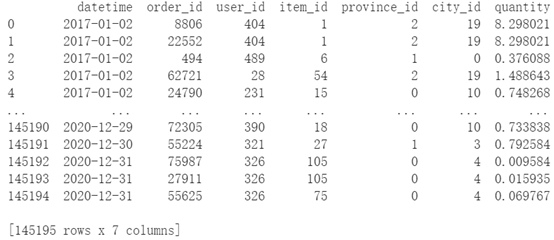
There is not much difference on the sale between each month. We can see that in the winter the sales are not as good as other seasons. The overall trend for years is decreasing. We find that even if we have Covid-19 in 2020, the overall sale did not decrease a lot from 2019.

### **Data Mining Tasks**

Wine is an alcoholic drink typically made from fermented grapes. Yeast consumes the sugar in the grapes and converts it to ethanol, carbon dioxide and heat. Different varieties of grapes and strains of yeasts are major factors in different styles of wine.

Now, the team received one wine company’s desensitization data, which masked an important message of business replaced by numbers. The company hopes to receive predicted data in order to understand its selling information in the next several months.

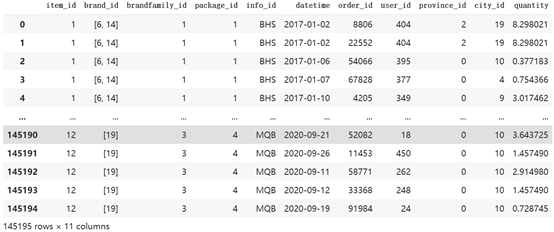
The team’s first objective is to analyze the sales performance of each product in each state, from 2017 to 2020, by statistical and visualization methods. And then the team will figure out how the sales performance changed in the past few years, especially in 2020 under the COVID-19 situation. Finally, the team will make a prediction for the future sales performance based on the past data.

There are two datasets that will be used in this project. The first one is the dataset for product information. There are 5 columns in this dataset. The column names are item id, brand id, brand family id, package id, and info id. And there are 119 rows for unique products.

The second dataset is the dataset for sales performance. There are 7 columns in this dataset. The column names are date, order id, user id, item id, state id, city id, and quantity. This dataset shows that in an unique order, how much quantity of a product is sold to a user, which stands for the liquid store, also the order date and the location information of this liquid store. There are 145196 rows in this dataset.

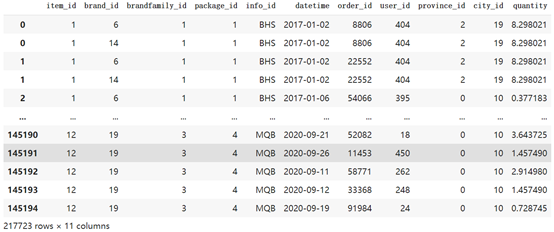
#### ***Data integrition***

There are two datasets, so the first step is to combine them together. After using pandas to read these two files, it shows that there is a common column called item id. Then these two tables can be joined together by the same item id. Also, set date as row key as it is not a feature.



After using pandas merge, the join result gets a table with 145195 rows and 11 columns, which is good.

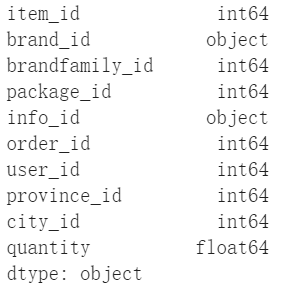
It can be seen that brand id has composite items as a list. In order to change its types to numeric, use the literal\_eval module from ast package to explode the list.



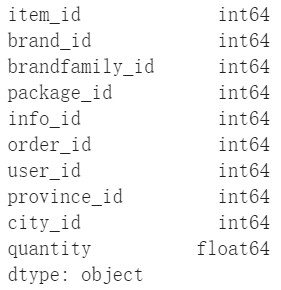
Which gives 217723 rows as a result.

#### ***Data Cleaning***

Since the dataset is the raw data from the company database, in order to clean the data, the first step is to check the data types.



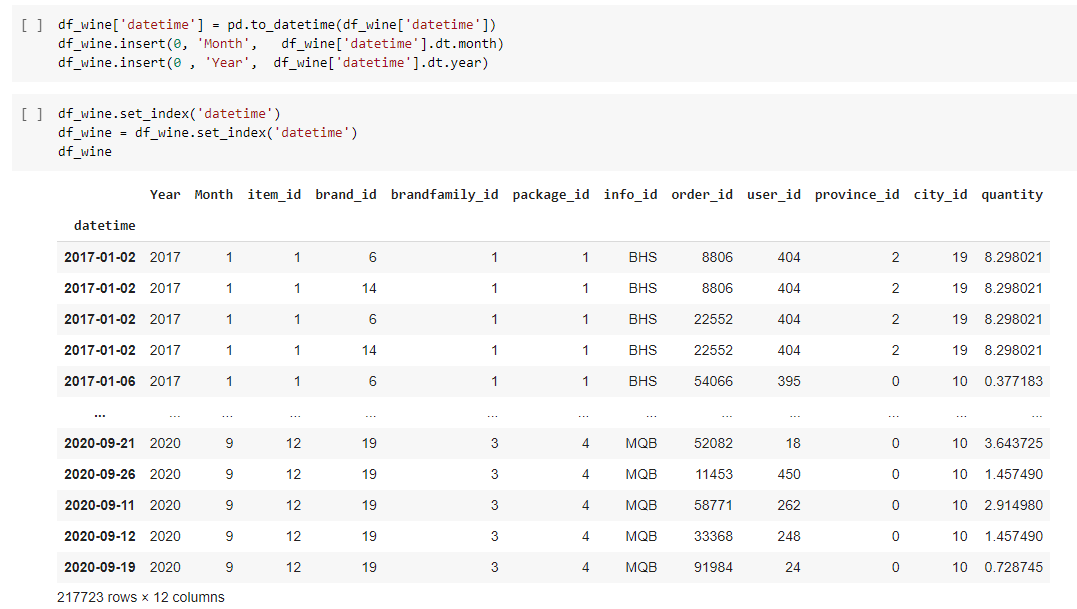
By using dtypes, it shows that the brand id and info id is an object, which is hard to handle. For the brand id, it can be converted to int64 by just using astype. But for info id, it should be converted to category and cat.codes it, then it can be converted to int 64 by astype.



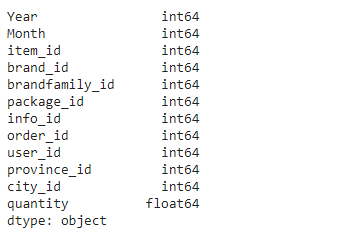
After converting the data type, it is necessary to check if there are any NA values by dropna or negative values in the dataset. Fortunately, there is no such a problem in this dataset.

#### ***Feature selections***

The team thinks it can not directly use datetime as the column as it has too many different possible values which are not proper for predictions. it is converted to datetime as an index, instead of adding ‘Year’ and ‘Month’ as new columns for predictions.



Check datatype again to make sure they consistent with other data



#### ***Prediction and Forecast***

The team will use machine learning methods to predict our target. There are 5 methods (KNN , Random Forest, Linear Regression, Neural Network, LightGBM) and they are all regression methods. The details of our models will be talked about below. The project uses previous years’ data to predict recent years’ data(all labels are available). After tuning parameters on all models, the team chooses the best model and predict on manually created test sets to forecast future values(both features and target are not available)

#### ***Hold Out Method***

#### The team will try Hold Out in the project

The data sets will be splitted into two parts at first. As the project is a forecasting project, the dataset was splitted by year. Data belonging to 2017,2018,2019 is in the training set, data belonging to 2020 is in validation. The train set is used on every model and the performance is evaluated on validation model

### **Data Mining Models/Methods**

#### ***KNN method***

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. For the certain data sample, it can find out the several train samples that are close to the train set based on the distance metrics. When the team preprocesses the dataset, it shows a good linear shape, so a KNN regression model would naturally fit this data well.

In order to find the value k to decide how many samples are close to the train data, the data standardization needs to be done firstly. Then the team decided to use normalized KNN with grid search to find the best performance model.

#### ***Random Forest Method***

Random Forest method is one of the most powerful machine learning methods that is used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. Every tree in the random forest is training a subset of the data.The basic idea behind it is to combine multiple decision trees in determining the final output, rather than relying on individual decision trees.Each decision tree has a high variance, but when we combine all decision trees together in parallel, the result variance is very low because each decision tree has been perfectly trained for specific sample data, so the output does not depend on one decision tree but on multiple decision trees. The final output is the average of all the decision tree outputs. Therefore, it can be seen as the fittest model for the current problem.

To improve the performance of the model, tuning hyperparameters is a necessary step. When the team tests the tuning method for the random forest model, it finds that even though the grid search is the better approach, it takes too much time to run the computation. Based on that situation, the team still chooses to find the best estimator: n and also the best max\_depth: d.

#### ***Multiple Linear Regression Method***

Multiple Linear Regression is the most simple and straightforward way for solving a regression problem with multiple predictors. The team will try three different types of linear regression: starting with normal linear regression and trying regularizations on linear regression: lasso regression(regularization with 1st norm) and ridge regression(regularization with 2rd norm). Applying gridsearch on the datasets will give more valid results for this method. The team will try to turn on and turn off “fit-intercept”and “normalize” for all our three linear models. Besides tuning those bool parameters, for lasso and ridge, by trying to tune the alpha values, it can be seen which alpha will generate better scores.

#### ***Neural Network Method***

Neural Network is one of the most popular and powerful models nowadays. It usually takes a long time to run but generally gives good results as data is handled in detail in hidden layers. The data was scaled at first as it was already known that normalization will give better performance for neural networks generally. Then we tune our parameters. Here we try 4 combinations( 0.001, 0.01, 0.1, 1) of learning rate and 3 combinations of transfer functions（'identity', 'logistic', 'relu').

#### ***LightGBM Method***

LightGBM is a state of the art model and has been very popular in industry in recent years. As it is a model not that familiar compared with other models and also it takes a long time to run. The parameter tuning processing will not be complicated. Num\_leaves is increased once (64 to 100) and max\_bin is decreased once (255 to 100) to compare performances

### **Performance Evaluation**

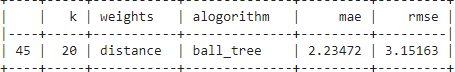
As the problem is a regression problem and also forecast problem, the team uses MAE, and RMSE, together with comparing forecast values with actual values in a line graph to evaluate the model performance. Those are regression metrics as our metrics to evaluate our models through the whole project. RMSE is chosen as the main metric to compare and identify which model is better, as it is able to give a fair evaluation about our loss in regression models. The line graph is used to see how close the predicted line and actual are fitting each other.

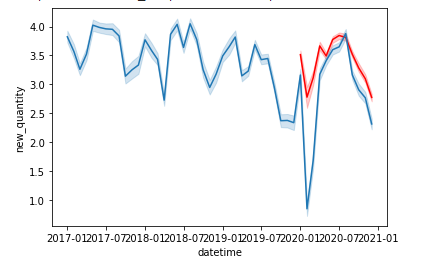
### **Project Results**

#### ***Best Model Selection***

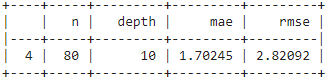
We check performance by comparing both RMSE and closeness of predicted line and actual line.

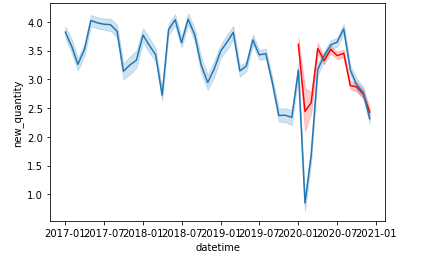
***KNN:***



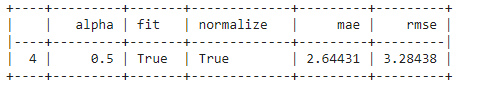


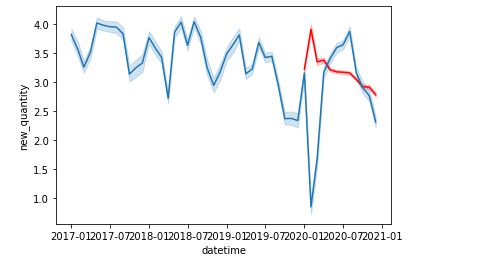
***Random Forest:***

******

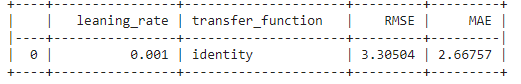
******

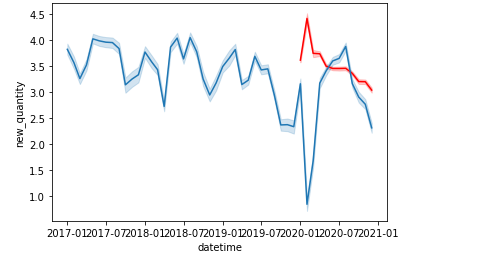
***Linear Regression:***

******

******

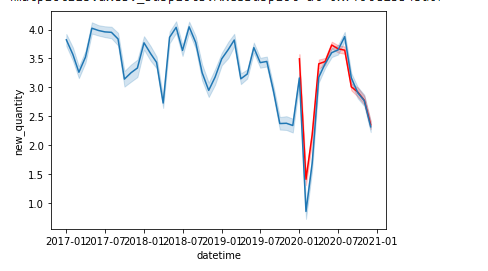
***Nerual Network:***

******

******

***LightGBM:***

******

******

Rank the performance from best to worst:

**RMSE:**

LightGBM(2.18)

Random Forest(2.82)

KNN(3.15)

Linear Regression(3.28)

Neural Network(3.30)

**Predicted Line:**

LightGBM

Random Forest

KNN

Linear Regression

Neural Network

**Overall Performance:**

LightGBM

Random Forest

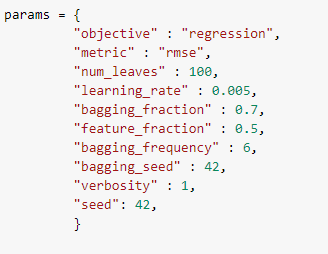
KNN

Linear Regression

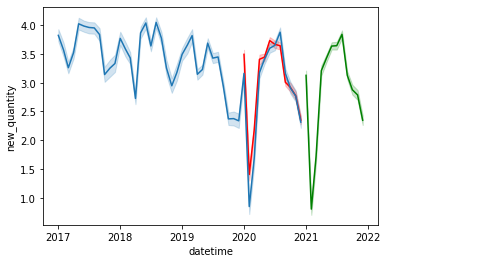
Neural Network

#### ***Predict on Test sets***

The team chooses the best model: LightGBM with below parameters:



As we lack an available data set for next year as test data , manually made data was necessary. The data distribution for year 2021 was assumed to be the same as data in 2020, The only difference is the year was changed to 2021. The train set will also cover more data to increase the predicting accuracy, the most recent year 2020 will be included in the training phase instead of in the validation phase. In other words, the best model was trained on the whole dataset and test on the manually made fake data to predict next year’s results.



The trend of year 2021 and 2020 was very similar as the data of 2021 mocks data of 2020. The forecasting is successful but this situation is not likely to happen in real life as the trend was too similar. After training on existing data, how to make predictions about future results based on available existing data is a tough question and should be considered as a significant problem next time.

### **Impact of the Project Outcomes**

With the final prediction based on the wine company’s original selling data, it will provide its predicted data to draw the next several month's selling needs. By looking at the predictive data structure, the wine company can target certain hot products and reduce production of certain products to avoid unnecessary losses.In addition, in view of different forecast sales in different regions, the company can deploy transportation capacity and the Sales Department in advance to connect with local dealers, and the Marketing Department can also make further work arrangements for regions with weak forecast sales. While Covid-19 affects the accuracy of predictive models, it does provide a high degree of insight into how the company will perform in the coming months.