Submitted By: Anish Bhusal
March 18,2020

Linear Regression Using Kalimati Market Price Dataset

1. Data Analysis and Visualization

The dataset consists of daily vegetable prices obtained from Kalimati Market Price sheet daily from 2018-2020 Feb. It consists of recorded maximum, minimum and average prices for vegetables with respect to wholesale and retail market.

Total Records	99302
Total Unique Veggies	120

For easier analysis, the veggies were grouped into three categories:

Fruits	Vegetables	Meat

The prices of fruits, vegetables and meat showed following distribution in the market:

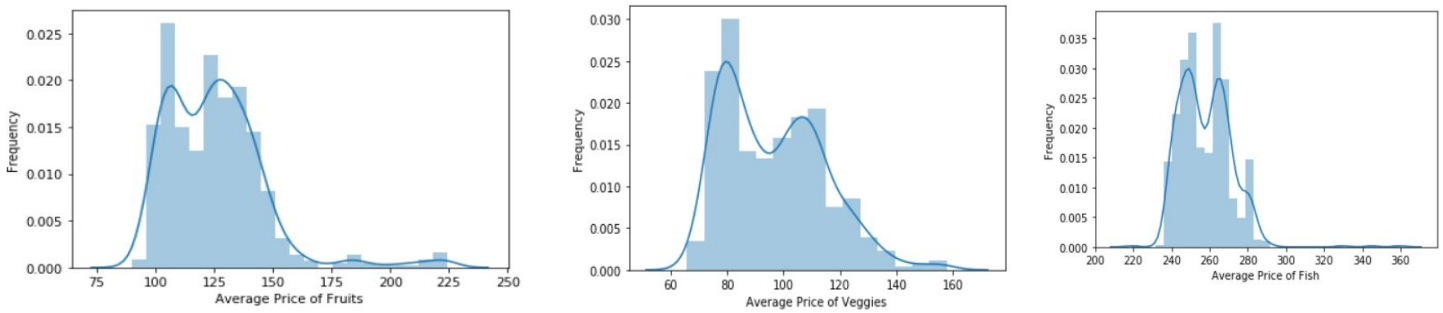
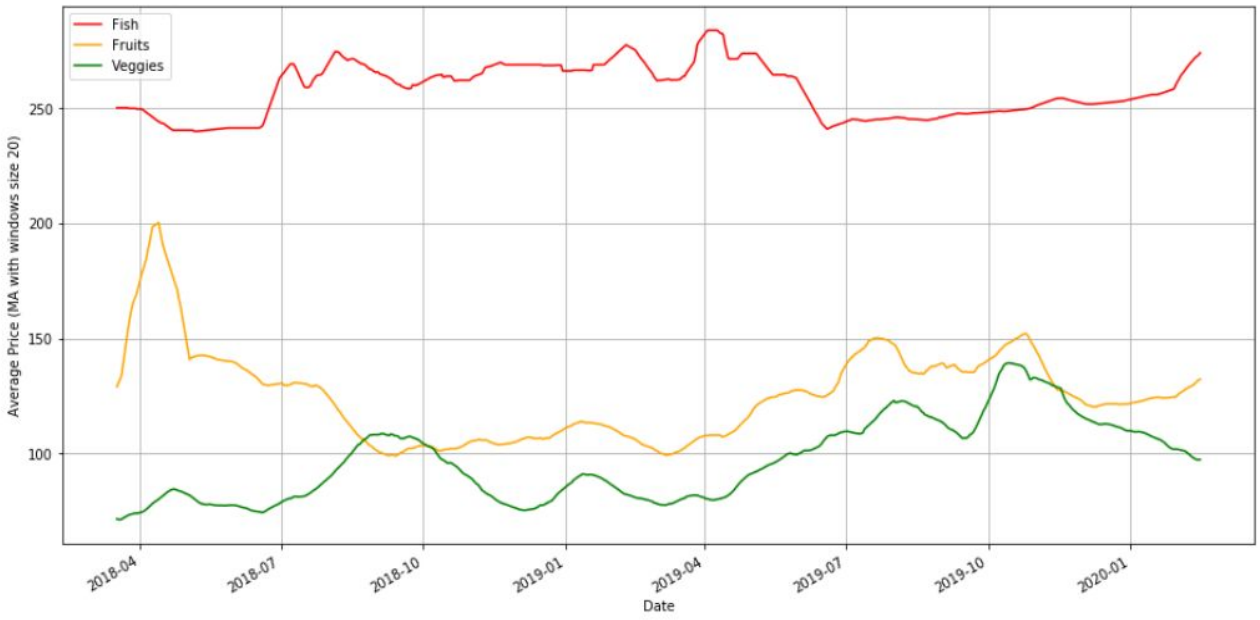


Fig: Distribution Plot for Fruits, Vegetables and Fish(Meat)

The average price of fruits mostly lies between 100-150 price range, for vegetables it is between 70-90 and for fish(meat) it's between 250-270. The following graph shows how prices varied over time for these three categories:



Submitted By: Anish Bhusal
March 18,2020

2. Feature Extraction and Feature Normalization

Following features were chosen for training:

Veggies | Day - 1 Price | Day -2 Price | Price Type

The dataset was then split into train set, validation and test set into 60:20:20 ratio.
Min-Max normalization was carried out in features.

3. Model Training, Grid Search and Evaluation

Model was trained, evaluated and grid search was carried out with following params:

num_of_iterations	[1000,1500,2000,2500,3000]
learning_rate	[0.001,0.1,0.01]
Lambda_l2 (L2 Regularization)	[100,200,300]
Loss Function	MSE Loss
Model Evaluation Metric	R2 Score

The best model with least loss was selected with following: hyperparameters

num_of_iterations	3000
learning_rate	0.1
lambda_l2	300
MSE Loss	6193.804

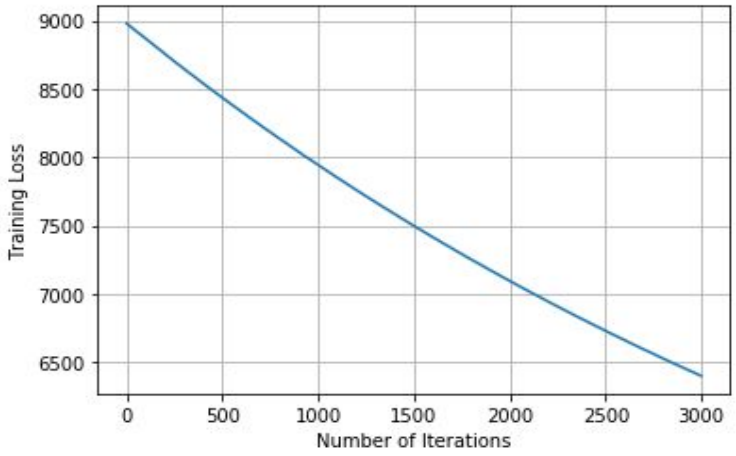


Fig.: Training Loss vs Number of Iterations

R2-Score for selected model:

R2-Score	0.320955

R2 Score is greater than 0 for our predictions which is good. This model captures some of the variability of the data around its mean but more work is necessary to capture almost all of the variability. Adding other feature columns like seasonal information (which affect veggies price a lot), local demand, etc could help to predict prices in more accurate way.

Code for this analysis is available here.