Predicting Canadian Consumer Food Prices with Python

Patrick Walter

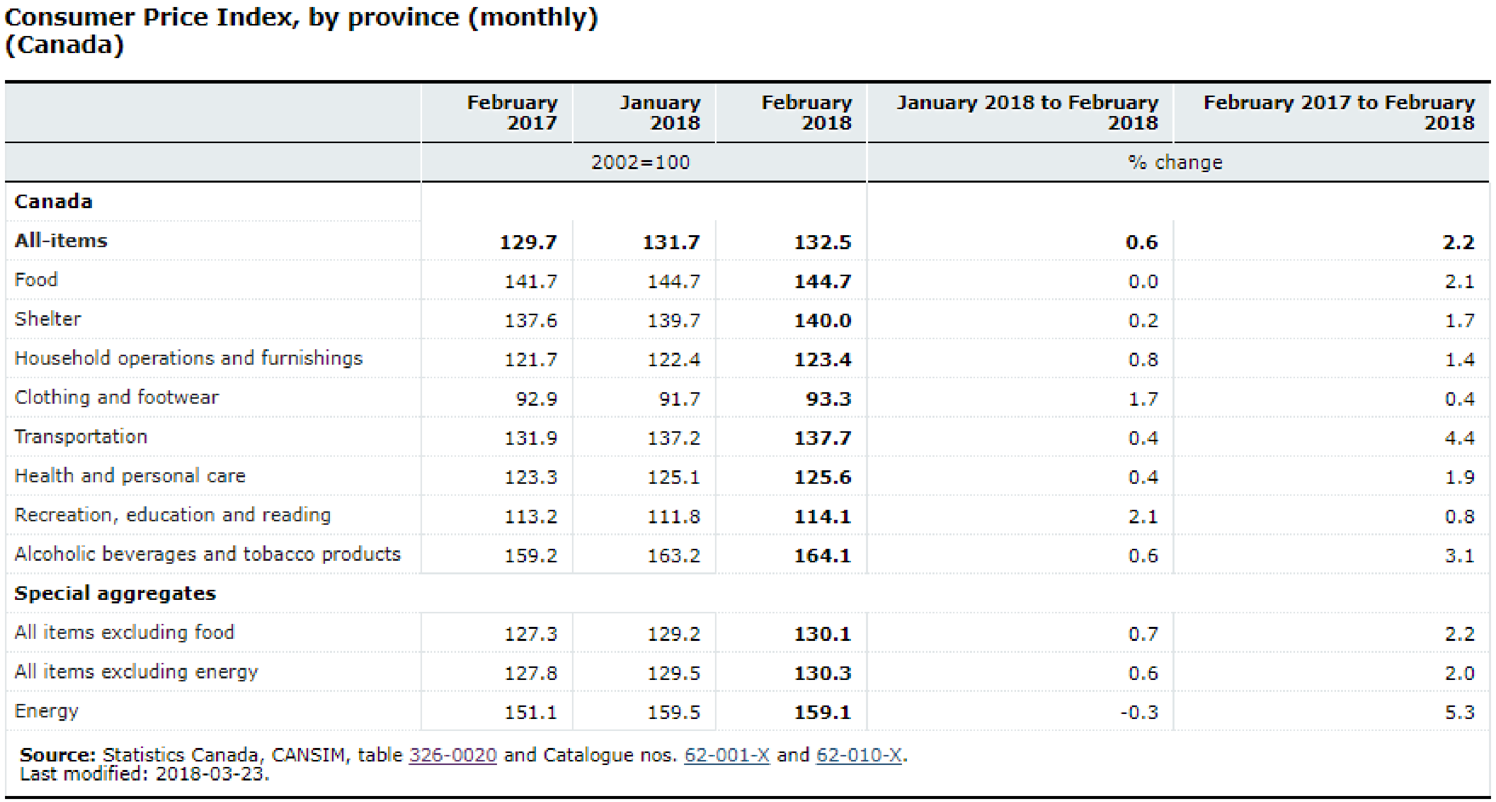
Background: Canada's Food Price Report

- This year the 8th edition of Canada's Food Price Report was published as a collaborative effort between Dalhousie University and University of Guelph.
- The report looks at factors that affect the future prices seen by consumers for food over the next 12-month period.
- Includes the key drivers of the upcoming year's food prices such as Climate, Energy Costs, Inflation, Policy Context, Food Processing Industry, Consumer Debt and Deleveraging and many more.
- The report divides the CPI basket into food categories including Dairy and Eggs, Fruit and Nuts, Meats, Vegetables, and Fish and Seafood.

Background: Consumer Price Index

- An indicator of the changes in prices experienced by Canadians, obtained by comparing the rising or falling cost of a fixed basket of good and services.
- Fixed basket means that products and services are of the same quality and quantity and therefore these changes in cost are a true reflection of the pure price changes.
- Data is released monthly and open to the public on Statistics Canada’s website as part of the CANSIM database.

Background: Consumer Price Index



Consumer Price Index, by province (monthly) (Canada)

	February 2017	January 2018	February 2018	January 2018 to February 2018	February 2017 to February 2018
	2002=100			% change	
Canada					
All-items	**129.7**	**131.7**	**132.5**	**0.6**	**2.2**
Food	141.7	144.7	**144.7**	0.0	2.1
Shelter	137.6	139.7	**140.0**	0.2	1.7
Household operations and furnishings	121.7	122.4	**123.4**	0.8	1.4
Clothing and footwear	92.9	91.7	**93.3**	1.7	0.4
Transportation	131.9	137.2	**137.7**	0.4	4.4
Health and personal care	123.3	125.1	**125.6**	0.4	1.9
Recreation, education and reading	113.2	111.8	**114.1**	2.1	0.8
Alcoholic beverages and tobacco products	159.2	163.2	**164.1**	0.6	3.1
Special aggregates					
All items excluding food	127.3	129.2	**130.1**	0.7	2.2
All items excluding energy	127.8	129.5	**130.3**	0.6	2.0
Energy	151.1	159.5	**159.1**	-0.3	5.3

Source: Statistics Canada, CANSIM, table 326-0020 and Catalogue nos. 62-001-X and 62-010-X.
Last modified: 2018-03-23.

Methodology: Research Objectives

- Train and test models for predicting the average national food price as reported by the CPI for a twelve month period for 5 years, from 2012 to 2016.
- Train and test predictive models for predicting 21 other targets in the CPI basket for the years 2016 and 2017.
- Gain insight on what years were difficult to predict, and what products were difficult to predict.
- Implement an ensemble method to improve the stability of the models to predict the average national food price for the twelve months reported by the CPI for 2016 and 2017
- Gain insight on how effective ensemble methods are at making stable predictions on average national food price.

Methodology: Dataset

- The original dataset was built by Jay Harris for his work on the 2017 and 2018 Canada Food Price Report and his thesis research titled *A Machine Learning Approach to Forecasting Consumer Food Price*
- It was constructed using a wide range of financial and econometric data from public and private institutions totaling 280 attributes
- The data set began at January 1985 with monthly records concluding at August 2017 giving 291 records.
- 22 targets from the CPI basket, mostly food products.

Methodology: Data Preprocessing

- Pandas *dataframes* was used to manipulate the data.
- *Dataframe* is a tabular data structure that has labeled axes.
- Dividing the dataset in targets and attributes
- Capping the dataset at 1999 after feature selection.
- Removing any columns that were incomplete
- Creating train and test *dataframes* for each of the five years
- Creating *dataframes* for each of the individual targets

Tools: Python Scikit Learn

- A machine learning library for Python with a wide variety of implementations for popular algorithms.
- This research used the ordinary least squares linear regression implementation. Easy to use to train and test linear regression models.
- Feature selection, using mutual information and a selector to select the highest scoring features.

Predicting Average Food Price: 2012 to 2016

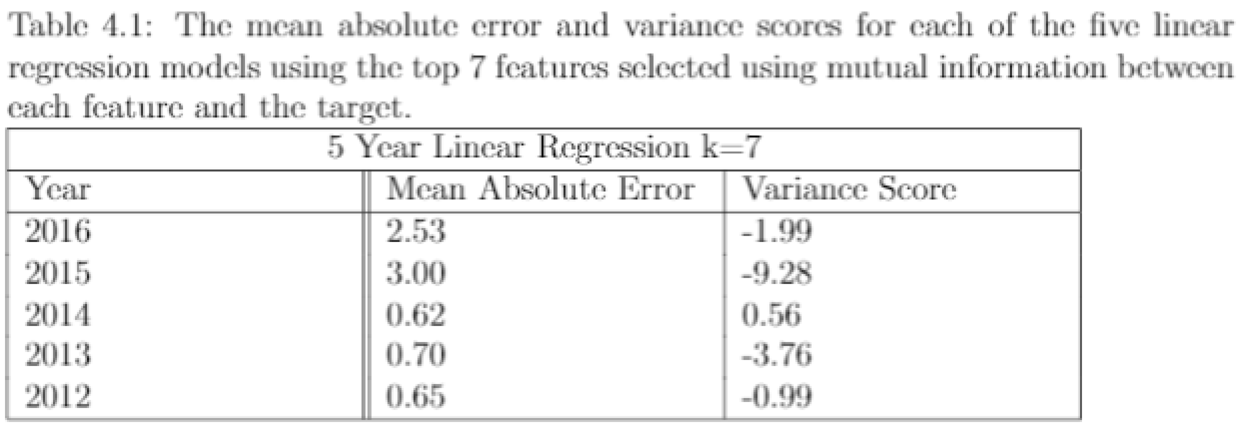
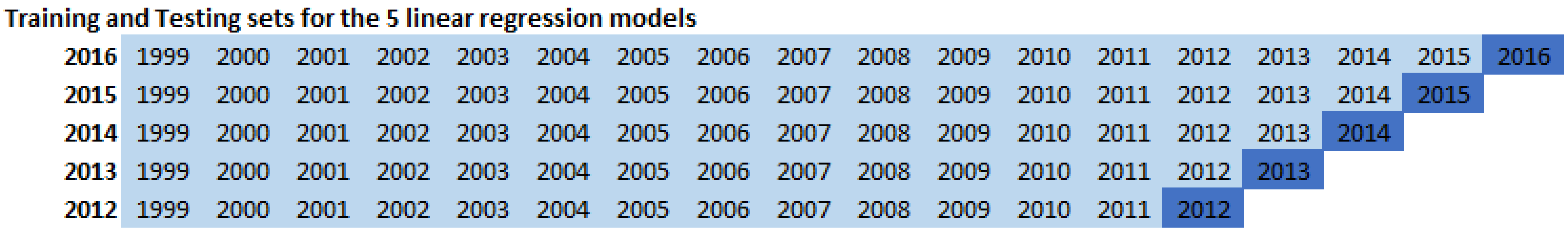


Table 4.1: The mean absolute error and variance scores for each of the five linear regression models using the top 7 features selected using mutual information between each feature and the target.

5 Year Linear Regression k=7		
Year	Mean Absolute Error	Variance Score
2016	2.53	-1.99
2015	3.00	-9.28
2014	0.62	0.56
2013	0.70	-3.76
2012	0.65	-0.99

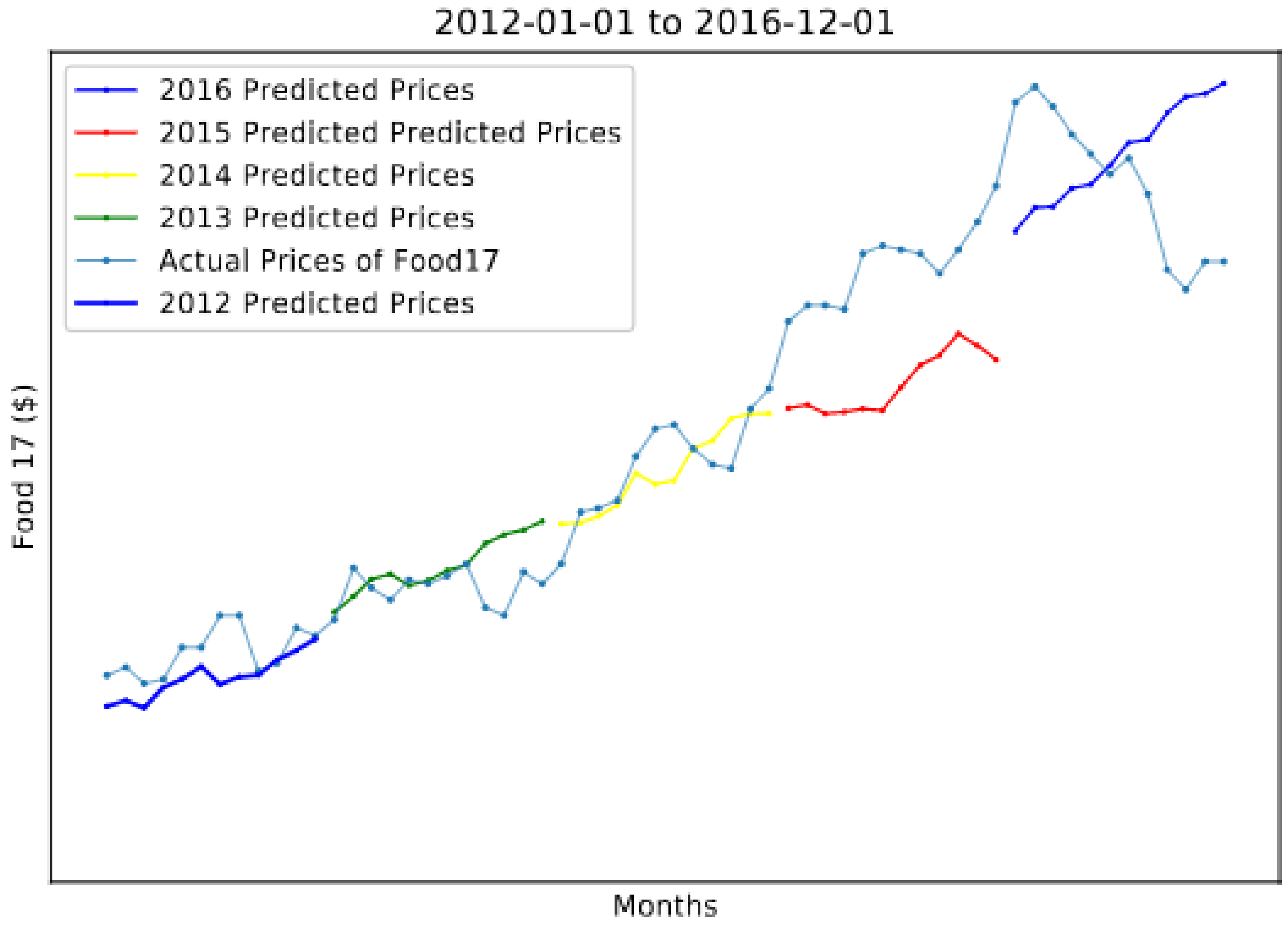
- Divided the data into training and testing sets for each of the 5 years
- Trained 5 individual models, tested each model on its corresponding year
- Mean absolute error was calculated across all 12 points for each of the 5 years.



Training and Testing sets for the 5 linear regression models

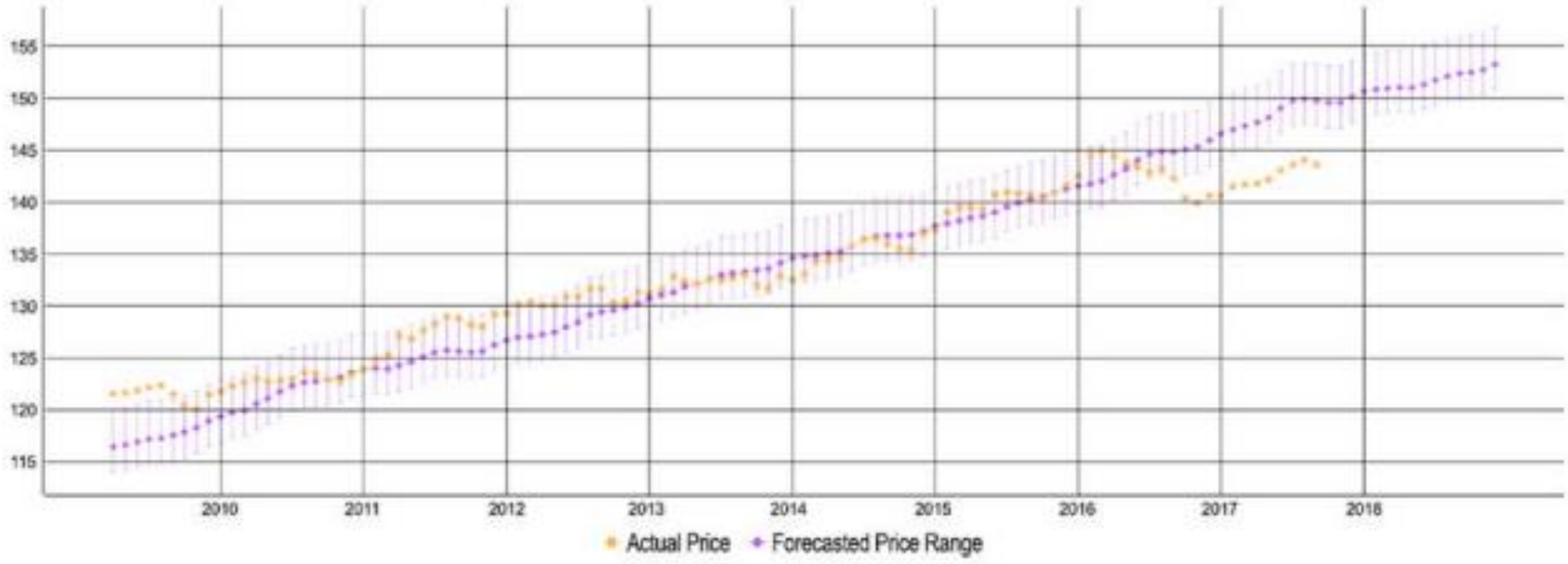
2016	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
2015	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	
2014	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014		
2013	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013			
2012	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012				

Predicting Average Food Price: 2012 to 2016



Food Price Report 2018

FIGURE 1: FOOD PRICES IN CANADA: ACTUAL VS. FORECAST



Predicting 22 Individual Targets:

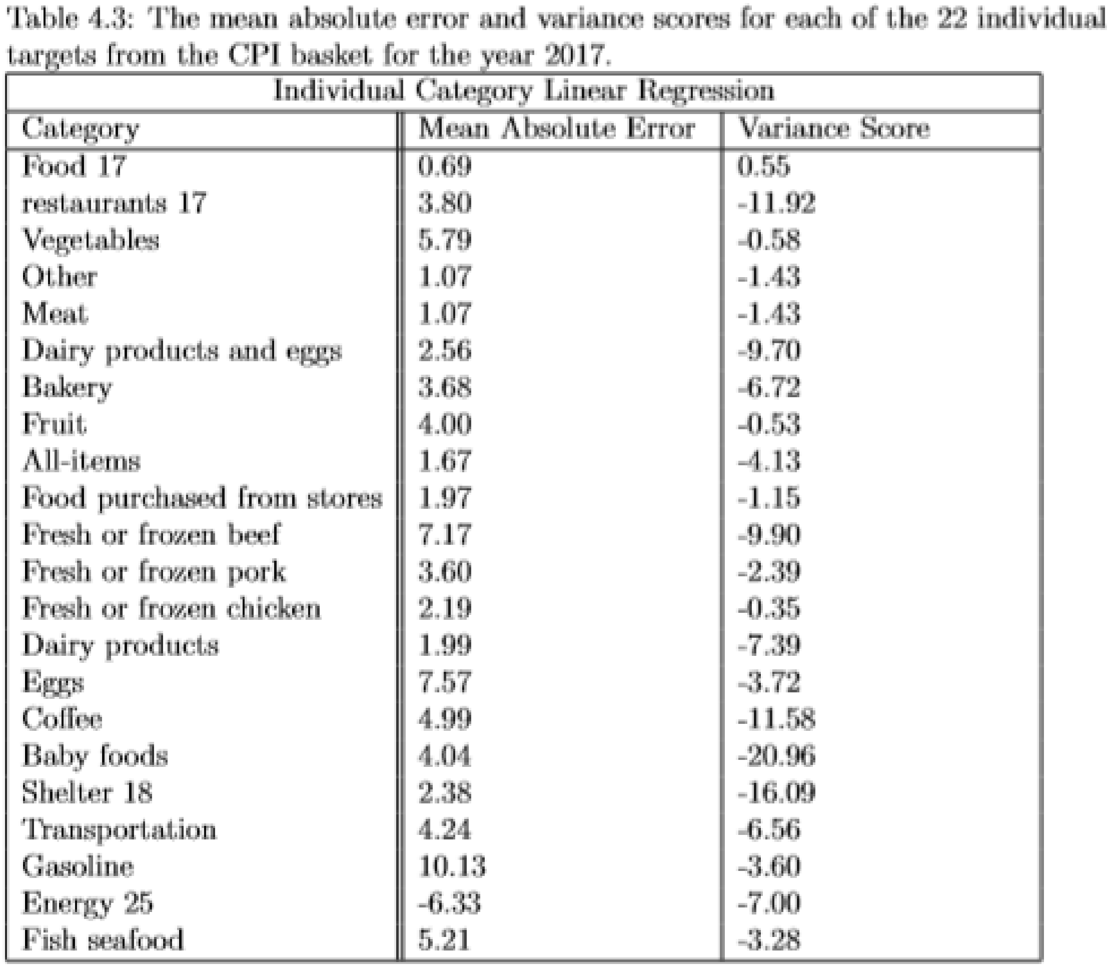


Table 4.3: The mean absolute error and variance scores for each of the 22 individual targets from the CPI basket for the year 2017.

Individual Category Linear Regression		
Category	Mean Absolute Error	Variance Score
Food 17	0.69	0.55
restaurants 17	3.80	-11.92
Vegetables	5.79	-0.58
Other	1.07	-1.43
Meat	1.07	-1.43
Dairy products and eggs	2.56	-9.70
Bakery	3.68	-6.72
Fruit	4.00	-0.53
All-items	1.67	-4.13
Food purchased from stores	1.97	-1.15
Fresh or frozen beef	7.17	-9.90
Fresh or frozen pork	3.60	-2.39
Fresh or frozen chicken	2.19	-0.35
Dairy products	1.99	-7.39
Eggs	7.57	-3.72
Coffee	4.99	-11.58
Baby foods	4.04	-20.96
Shelter 18	2.38	-16.09
Transportation	4.24	-6.56
Gasoline	10.13	-3.60
Energy 25	-6.33	-7.00
Fish seafood	5.21	-3.28

- 21 other targets were taken from the CPI basket
- Mostly products within the food category.
- Feature selection was done for each of the targets
- New models were trained and tested with each of the new targets with their selected feature.

Predicting 22 Individual Targets:

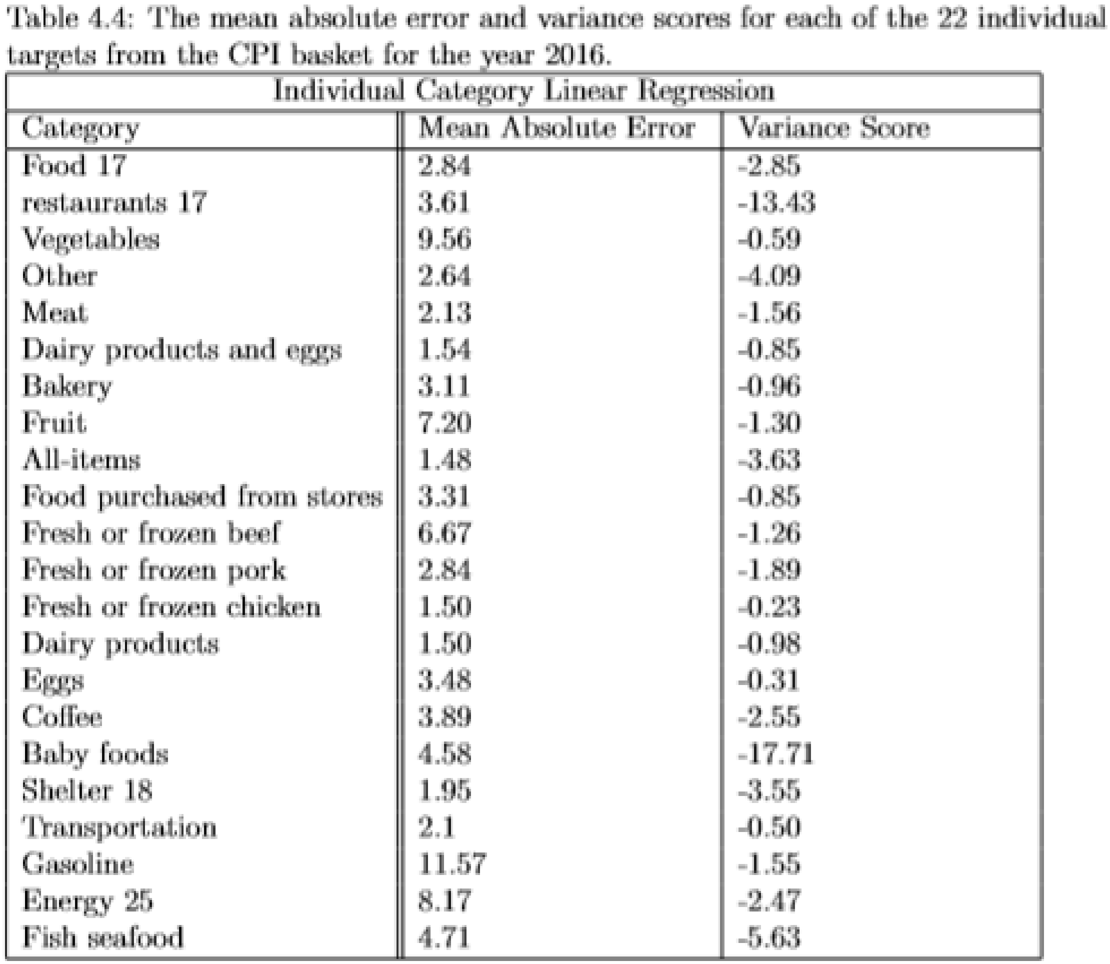


Table 4.4: The mean absolute error and variance scores for each of the 22 individual targets from the CPI basket for the year 2016.

Individual Category Linear Regression		
Category	Mean Absolute Error	Variance Score
Food 17	2.84	-2.85
restaurants 17	3.61	-13.43
Vegetables	9.56	-0.59
Other	2.64	-4.09
Meat	2.13	-1.56
Dairy products and eggs	1.54	-0.85
Bakery	3.11	-0.96
Fruit	7.20	-1.30
All-items	1.48	-3.63
Food purchased from stores	3.31	-0.85
Fresh or frozen beef	6.67	-1.26
Fresh or frozen pork	2.84	-1.89
Fresh or frozen chicken	1.50	-0.23
Dairy products	1.50	-0.98
Eggs	3.48	-0.31
Coffee	3.89	-2.55
Baby foods	4.58	-17.71
Shelter 18	1.95	-3.55
Transportation	2.1	-0.50
Gasoline	11.57	-1.55
Energy 25	8.17	-2.47
Fish seafood	4.71	-5.63

- 21 other targets were taken from the CPI basket
- Mostly products within the food category.
- Feature selection was done for each of the targets
- New models were trained and tested with each of the new targets with their selected feature.

Ensembles: Bootstrap Aggregating

- Normally for bootstrap aggregating (bagging) the training dataset is randomly sampled, with replacement, to create a set of training sets to train individual models.
- For a timeseries dataset, randomly sampling the training data cannot be applied.
- Need a different method of creating samples or partitions of the dataset to be used to train the individual models.

Bootstrap Aggregating: Sampling

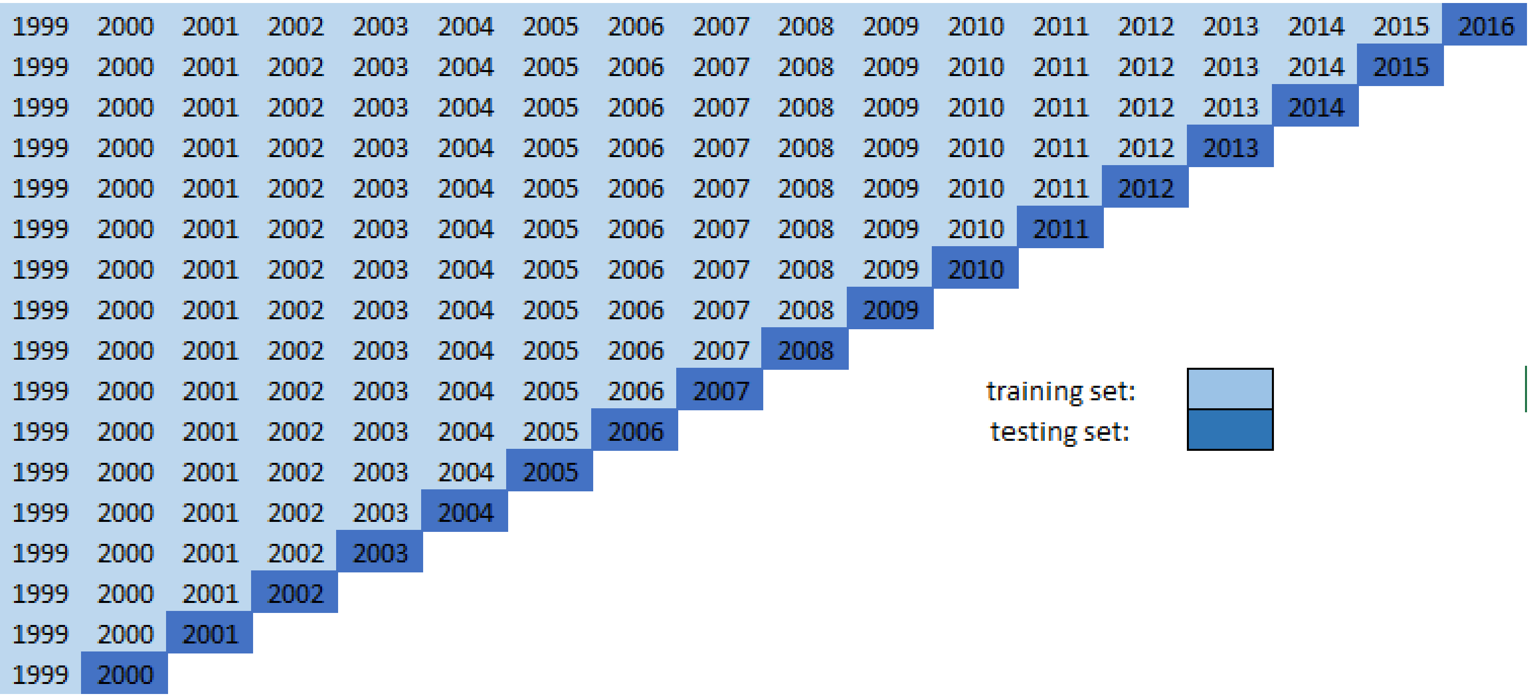
2015	2016
2014	2015
2013	2014
2012	2013
2011	2012
2010	2011
2009	2010
2008	2009
2007	2008
2006	2007
2005	2006
2004	2005
2003	2004
2002	2003
2001	2002
2000	2001
1999	2000

training set: ▭

testing set: ■

- First Idea was to simply partition the data by each year.
- Each model would only have 12 records to be trained on.
- Not enough examples to train stable models.

Bootstrap Aggregating: Sampling



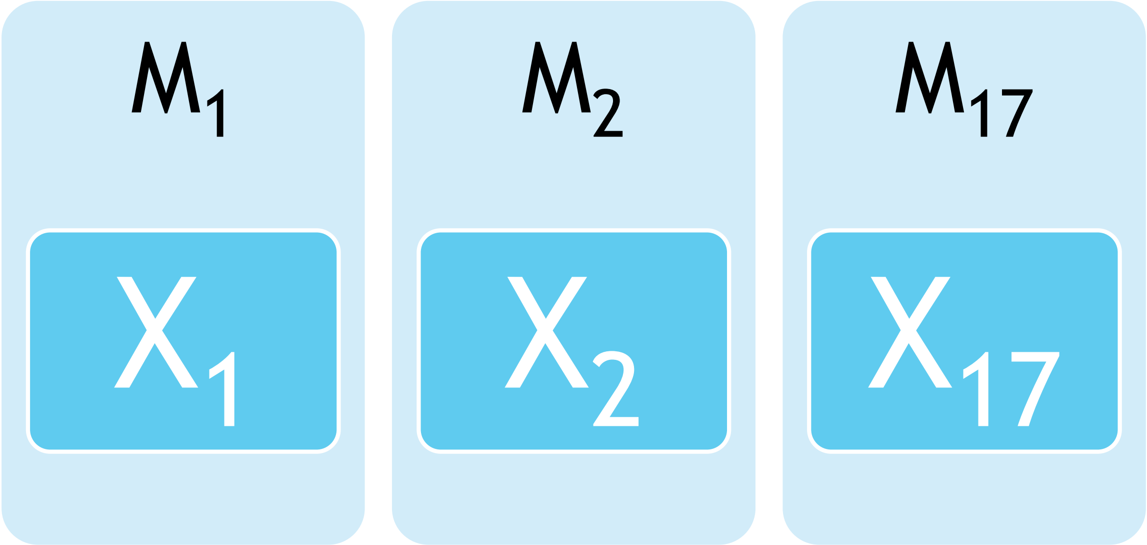
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014		
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013			
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012				
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011					
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010						
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009							
1999	2000	2001	2002	2003	2004	2005	2006	2007	2008								
1999	2000	2001	2002	2003	2004	2005	2006	2007									
1999	2000	2001	2002	2003	2004	2005	2006										
1999	2000	2001	2002	2003	2004	2005											
1999	2000	2001	2002	2003	2004												
1999	2000	2001	2002	2003													
1999	2000	2001	2002														
1999	2000	2001															
1999	2000																

training set: 🟦

testing set: 🟦

- The first sample contains January 1999 to December 2015.
- The second sample contains January 1999 to December 2014.
- …..
- The seventeenth sample contains January 1999 to December 2000.
- All together there are now 17 training sets, and 17 testing sets.

Bootstrap Aggregating: Training



M_1 X_1

M_2 X_2

M_{17} X_{17}

- Each model is trained and tested on it’s testing year.
- Each model is then trained and tested on one year ahead.

Bootstrap Aggregating: Testing

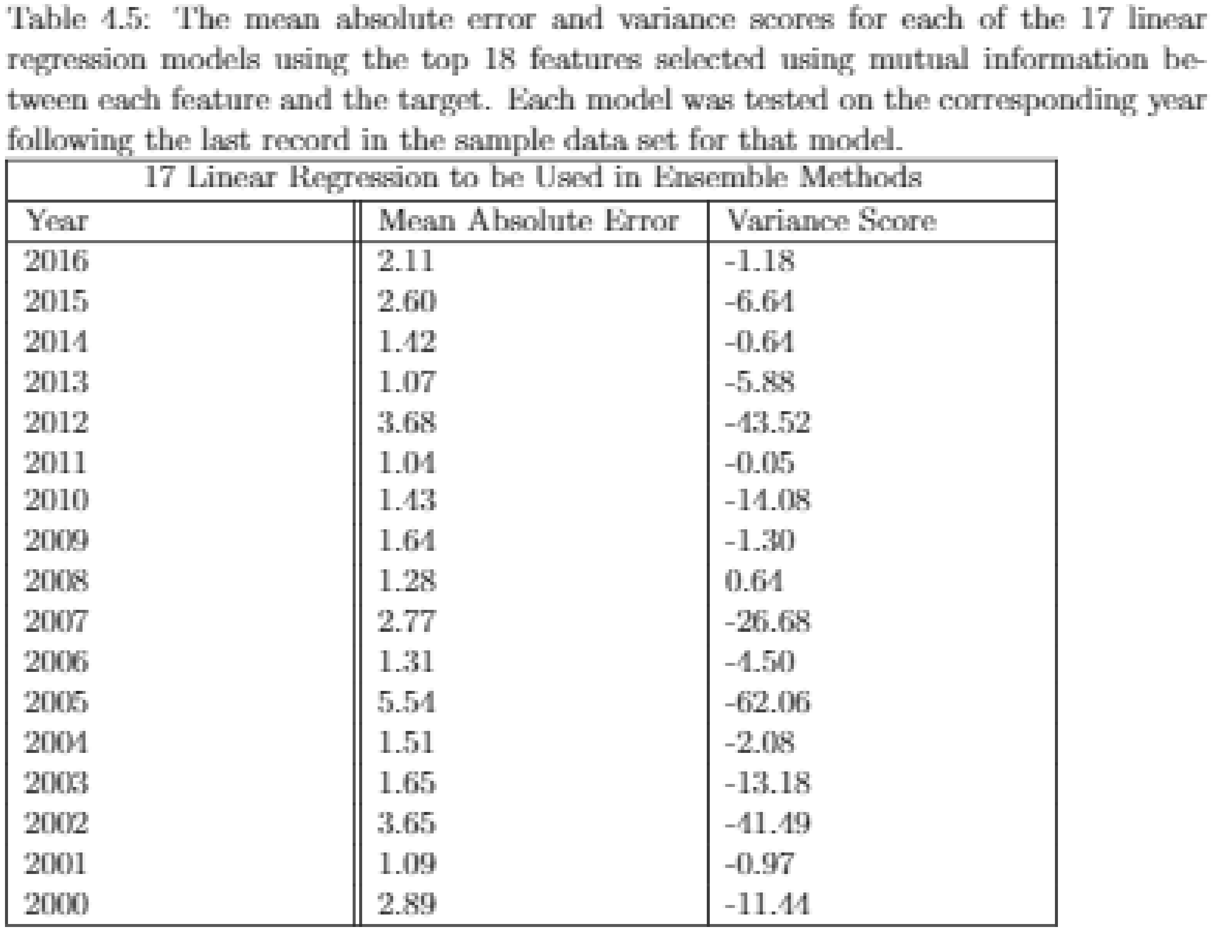


Table 4.5: The mean absolute error and variance scores for each of the 17 linear regression models using the top 18 features selected using mutual information between each feature and the target. Each model was tested on the corresponding year following the last record in the sample data set for that model.

17 Linear Regression to be Used in Ensemble Methods		
Year	Mean Absolute Error	Variance Score
2016	2.11	-1.18
2015	2.60	-6.64
2014	1.42	-0.64
2013	1.07	-5.88
2012	3.68	-43.52
2011	1.04	-0.05
2010	1.43	-14.08
2009	1.64	-1.30
2008	1.28	0.64
2007	2.77	-26.68
2006	1.31	-4.50
2005	5.54	-62.06
2004	1.51	-2.08
2003	1.65	-13.18
2002	3.65	-41.49
2001	1.09	-0.97
2000	2.89	-11.44

- Each model is trained and tested on it’s training year.
- Each model is then trained and tested on one year ahead.

Bootstrap Aggregating: Training and Testing

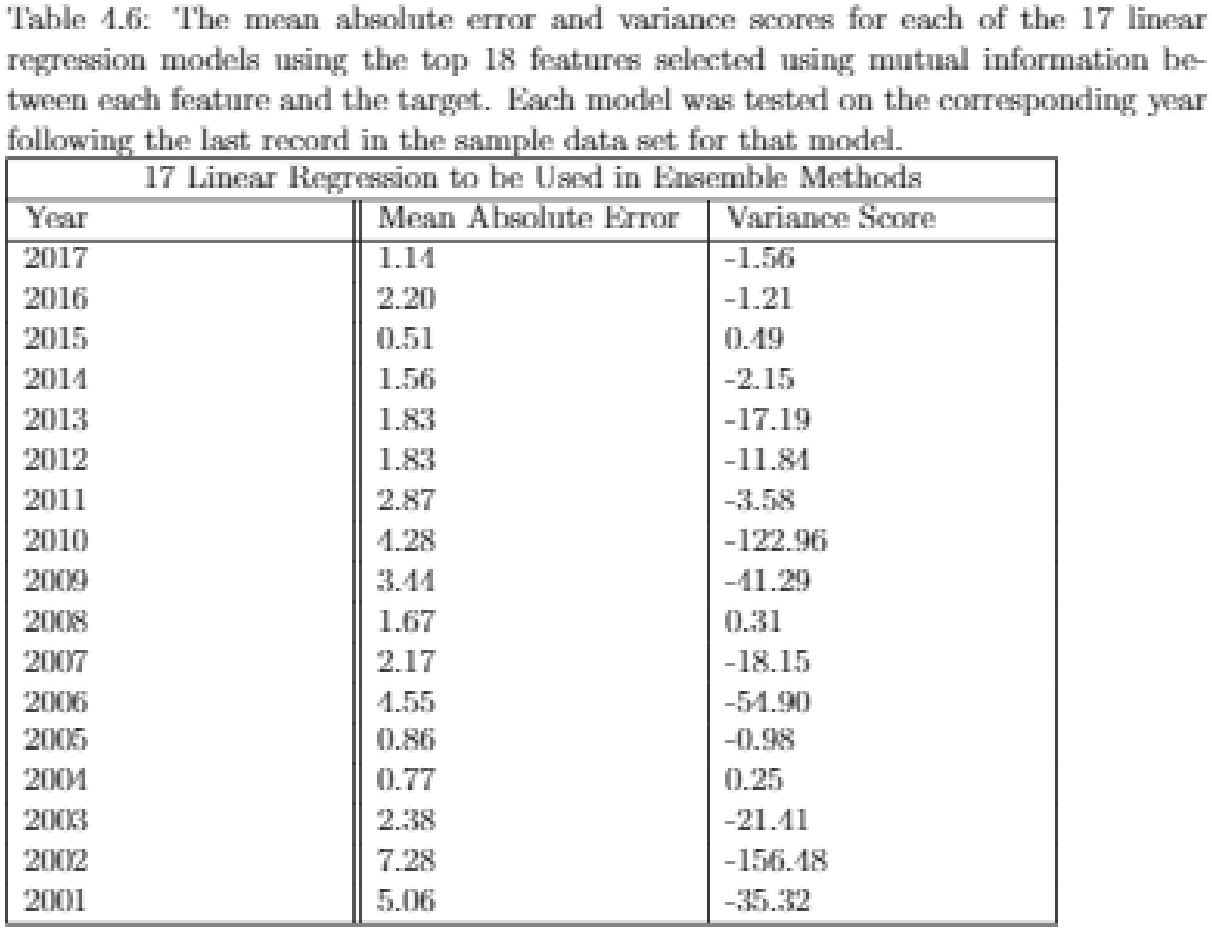
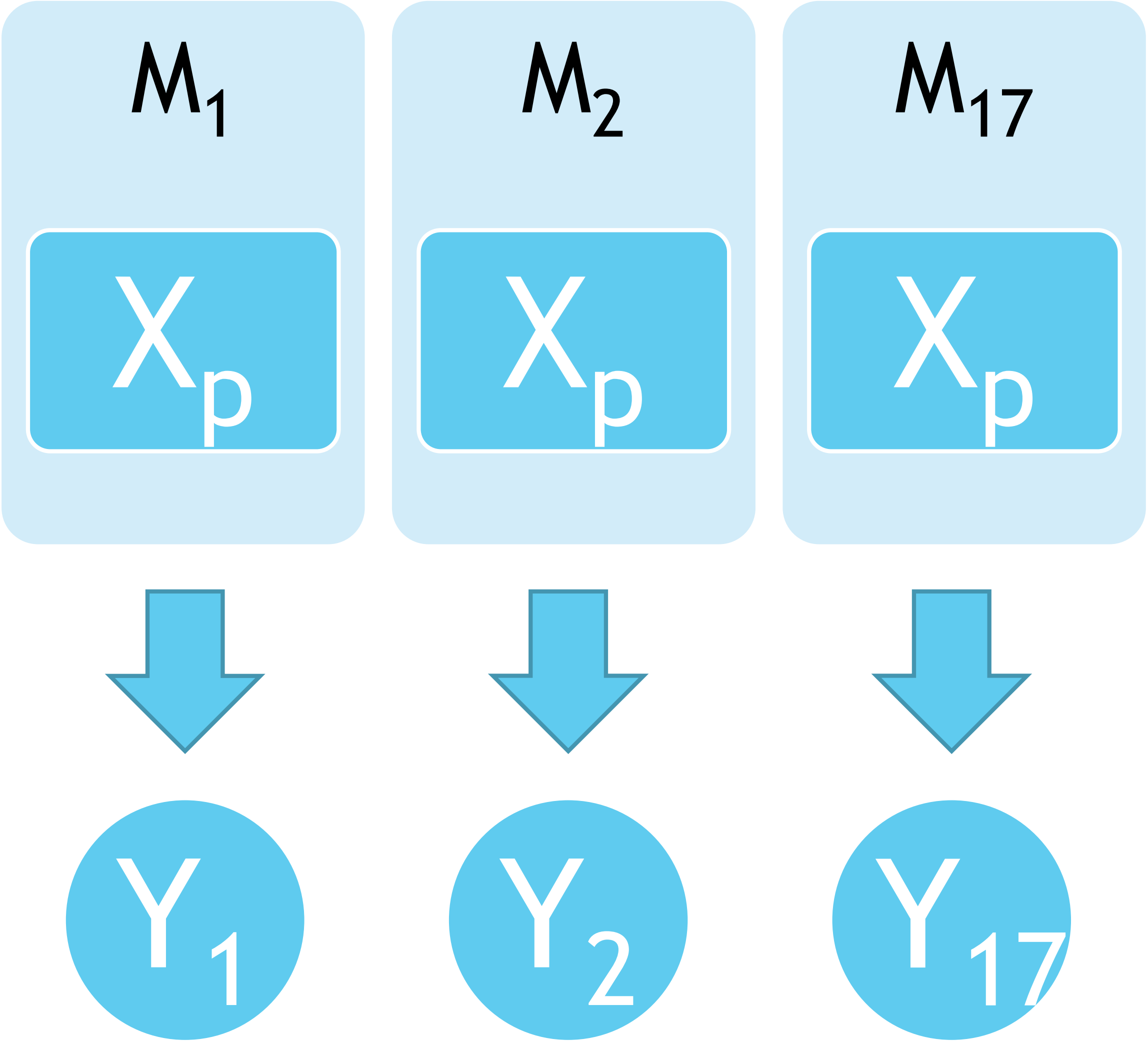


Table 4.6: The mean absolute error and variance scores for each of the 17 linear regression models using the top 18 features selected using mutual information between each feature and the target. Each model was tested on the corresponding year following the last record in the sample data set for that model.

17 Linear Regression to be Used in Ensemble Methods		
Year	Mean Absolute Error	Variance Score
2017	1.14	-1.56
2016	2.20	-1.21
2015	0.51	0.49
2014	1.56	-2.15
2013	1.83	-17.19
2012	1.83	-11.84
2011	2.87	-3.58
2010	4.28	-122.96
2009	3.44	-41.29
2008	1.67	0.31
2007	2.17	-18.15
2006	4.55	-54.90
2005	0.86	-0.98
2004	0.77	0.25
2003	2.38	-21.41
2002	7.28	-156.48
2001	5.06	-35.32

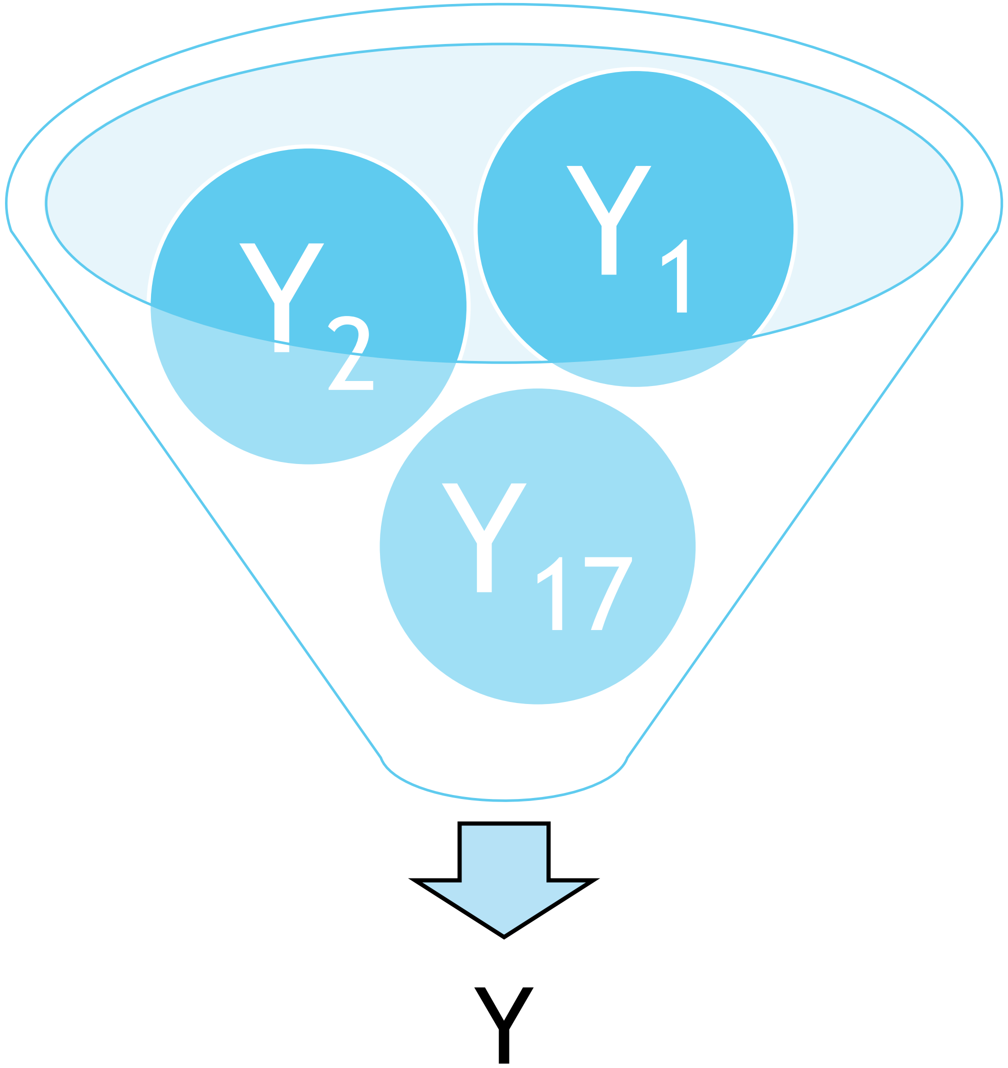
- Each model is trained and tested on it’s training year.
- Each model is then trained and tested on one year ahead.
- Results for predicting one year ahead were somewhat better.

Bootstrap Aggregating: Aggregating



- Use the target year's features in each of the models to produce predictions as output.
- For regression models we take the mean of the individual models as the aggregate output.
- Results for this were poor.
- Reduced to last 10 models
- Came up with a weighted average that gives decreasing weights to years further back.
- Next used a 1% inflation before taking the mean of the models.

Bootstrap Aggregating: Aggregating



- Use the target year's features in each of the models to produce predictions as output.
- For regression models we take the mean of the individual models as the aggregate output.
- Results for this were poor.
- Reduced to last 10 models
- Came up with a weighted average that gives decreasing weights to years further back.
- Next used a 1% inflation before taking the mean of the models.

Bootstrap Aggregating: Aggregating

$$w_i = 1/2^i$$

- Use the target year's features in each of the models to produce predictions as output.
- For regression models we take the mean of the individual models as the aggregate output.
- Results for this were poor.
- Reduced to last 10 models
- Came up with a weighted average that gives decreasing weights to years further back.
- Next used a 1% inflation before taking the mean of the models.

$$Y = Y_1 * w_1 + Y_2 * w_2 + Y_3 * w_3 ... + Y_{17} * w_{17}$$

Bootstrap Aggregating: Results

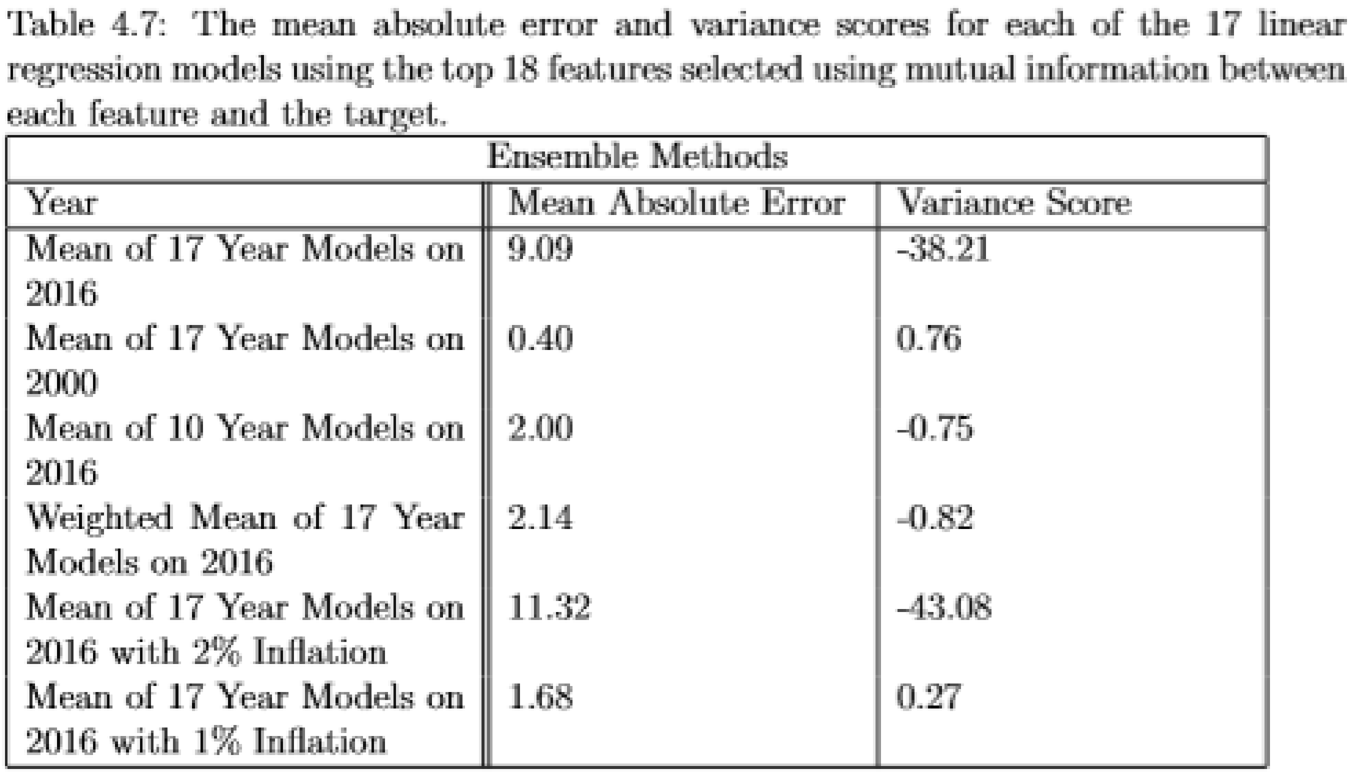


Table 4.7: The mean absolute error and variance scores for each of the 17 linear regression models using the top 18 features selected using mutual information between each feature and the target.

Ensemble Methods		
Year	Mean Absolute Error	Variance Score
Mean of 17 Year Models on 2016	9.09	-38.21
Mean of 17 Year Models on 2000	0.40	0.76
Mean of 10 Year Models on 2016	2.00	-0.75
Weighted Mean of 17 Year Models on 2016	2.14	-0.82
Mean of 17 Year Models on 2016 with 2% Inflation	11.32	-43.08
Mean of 17 Year Models on 2016 with 1% Inflation	1.68	0.27

- Each of the different bootstrap aggregating methods were trained and tested.
- The results were varied.
- Best predictive power for the year 2016 was inflation adjusted average.
- Both 10 year average and 17 year weighted average were significantly better than 17 year average.
- The normal single model for 2016 has MAE of 2.11

References

Charlebois, S., Jabez, H., Tyedmers, P., Bailey, M., Keselj, V., Conrad, C., . . . Chamberlain, S. (2017). Canada's Food Price Report. Halifax: Dalhousie University.

Harris, J. (2017, August). A Machine Learning Approach to Forecasting Consumer Food Prices. Halifax, Nova Scotia, Canada.

Hunter, J., Dale, D., Firing, E., & Droettboom, M. (2012). Matplotlib. Retrieved from Matplotlib: https://matplotlib.org/

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating Mutual Information. Physical Review, E(69).

Pandas. (2017, December). Python Data Analysis Library. Retrieved from Pandas: https://pandas.pydata.org/

Scikit Learn. (2017). Generalized Linear Models. Retrieved from Scikit Learn: http://scikitlearn.org/stable/modules/linear_model.html

Scikit Learn. (2017). sklearn.feature_selection.mutual_info_regression. Retrieved from Scikit Learn: http://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html#sklearn.feature_selection.mutual_info_regression

Statistics Canada. (2018, March). Consumer Price Index (CPI). Retrieved from Statistics Cnada: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2301