Group: Cormac O'Sullivan, Xinhui Luo, Weilin Cheng, Ishita Dutta, Rafael Garcia-Carreno

Contributions to Type II Diabetes

Introduction

Why do we want to analyze diabetes?

According to the CDC, Type II Diabetes affects 31 million Americans today; that is about 1 in 10 people. Type II diabetes causes the body to become insulin resistant. Insulin is a hormone produced by the pancreas that assists in letting sugar into cells. However, this resistance to insulin causes our cells to be unresponsive to insulin, thus leaving sugar in the blood and increasing blood sugar. High blood sugar causes all sorts of issues for the body including heart disease, kidney disease, vision loss, and restricted blood flow to appendages. The high amount of people that are inflicted with type II diabetes makes this an enormous health crisis in America today.

Our group is studying type II diabetes because a poor understanding of diabetes leads to worse diabetes outcomes. Education is the key to the prevention of this disease. In this project, our group will identify who is affected by type II diabetes and break down what groups in particular according to income, age, race, education level, and sex are most at risk. Then we will investigate further with a dataset containing more variables and see which one weighs more in our prediction.

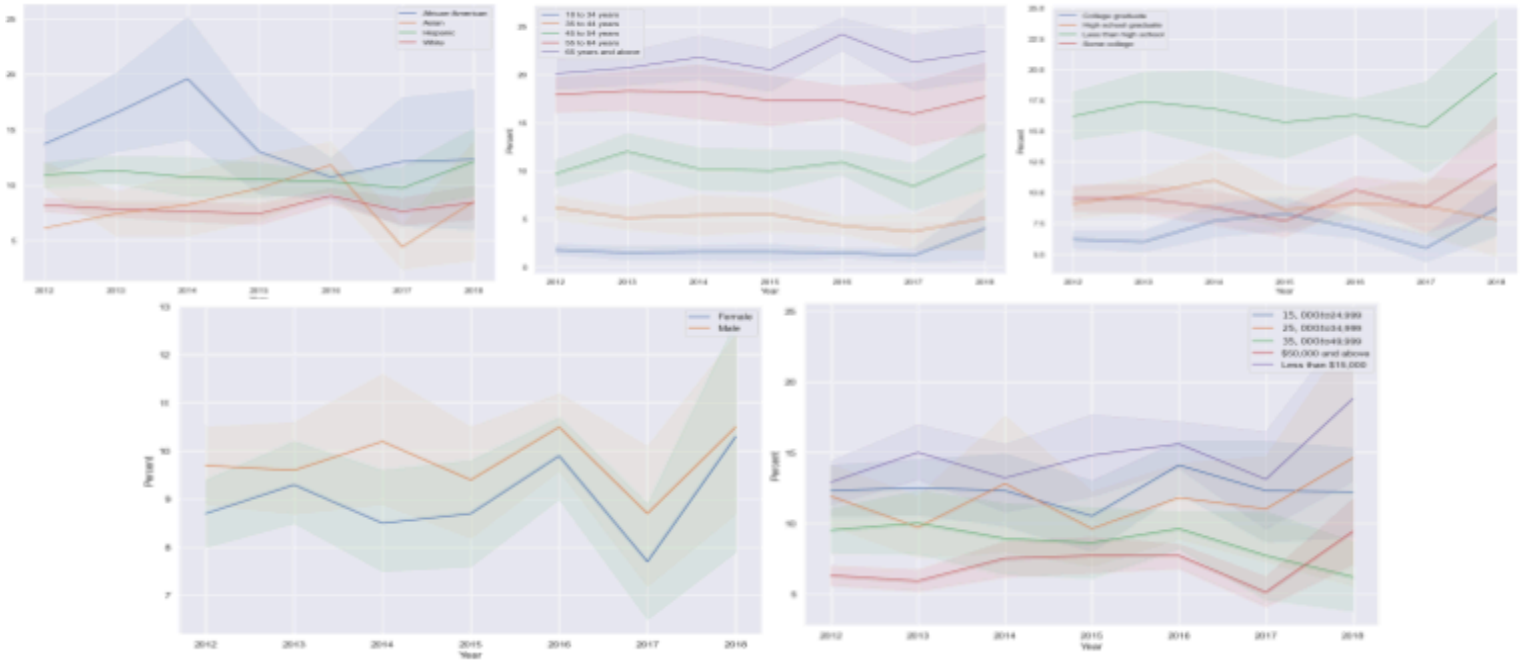
Many people in the United States are unable to get the necessary nutrients out of their food for a multitude of reasons; one being income. Here, our group will study the relationship between the price of bread, one of the most common staples in the American diet, and the quality of the nutritional content within them. This is our attempt to understand if an item is more unhealthy, then it is cheaper in American grocery stores.

Part 1 - Who is at risk?

We used the API method to get data from CA.GOV: Adults with Diabetes per 100 from 2012 to 2018. The data is from the California Behavioral Risk Factor Surveillance Survey (BRFSS). This dataset helps us gather general background information by providing us with a dataset consisting of percentages. It contains the following indicators:

a. **Race**: White, African-American, Asian, Hispanic
b. **Age**: 18 to 34 years, 35 to 44 years, 45 to 54 years, 55 to 64 years, 65 years and above
c. **Education**: Less than high school, High school graduate, Some college, College Graduate
d. **Income**: Less than $15,000, $15,000 to $24,999, $25,000 to $34,999, $35,000 to $49,999, $50,000 and above
e. **Sex**: male, female

We plotted each indicator to see how their trend changed over different years and compared each indicator within their strata group to see which group of people pay more attention to avoiding diabetes in California.



The **figures** above start from the top left: Race, age, education level, income, and sex.

The plots above are the rates of diabetes in California since 2012. While the plots alone are unable to give us any real predictive power, we do see some rough trends in the data that lend us some clues on what parameters of interest to potentially include in our model. In the age plot, we see the older group has a higher diabetes percentage. In the education plot, we see the group with an education level less than high school has a distinct difference compared to other groups. In the income plot, we see the group with a higher wage has less percentage of diabetes.
However, we do not want to make any conclusions just based on these plots since we did not perform any statistical test on this data. It only provides us with some ideas for the following parts. We know where to start our search for a good model and which factors to potentially look out for.

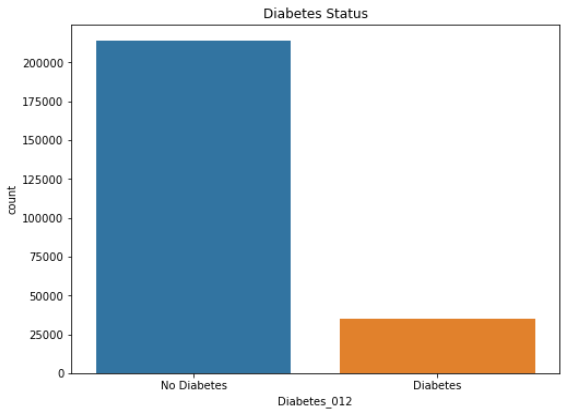
Part 2 - a national and more detailed analysis of indicators of diabetes

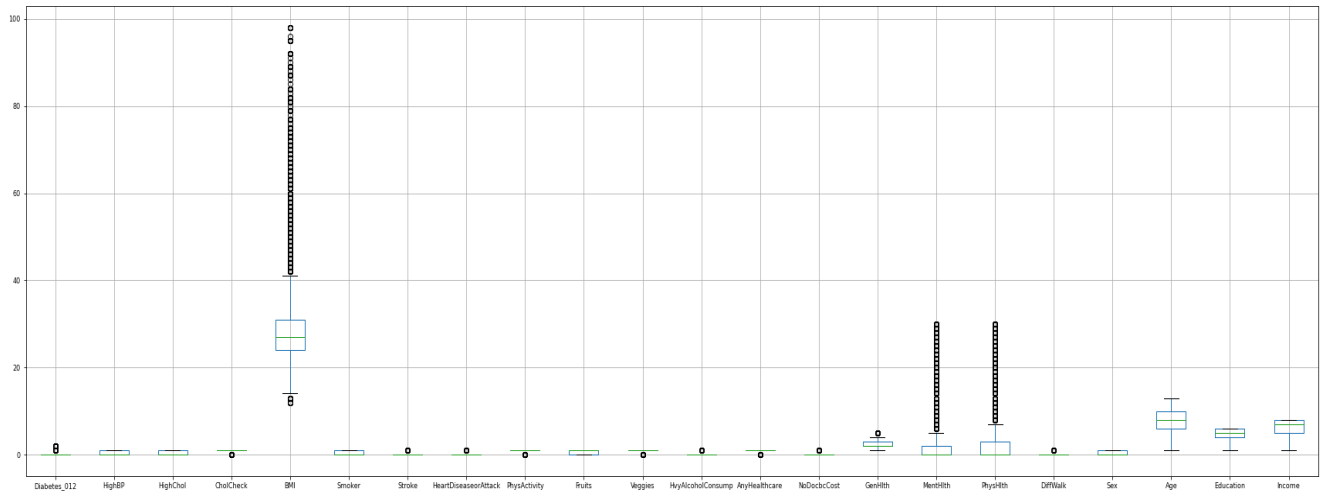
After the investigation in Part 1, we wondered if the whole US has a similar situation as California, so we found another dataset. This dataset comes from 253,680 survey responses to the CDC's BRFSS2015. It is a nationwide survey performed by the CDC in 2015.

Sex: 0 = female 1 = male	**HighBP:** 0 = no high BP 1 = high BP	**HighChol**: 0 = no high cholesterol 1 = high cholesterol	**BMI**: Body Mass Index	**Age:** 13-level age category: 1 = 18-24 9 = 60-64 13 = 80 +
GenHlth: What is your personal general health? 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor	**DiffWalk:** serious difficulty walking or climbing stairs 0 = no 1 = yes	**Fruits:** Consume Fruit 1 or more times per day 0 = no 1 = yes	**AnyHealthcare:** Do you have healthcare coverage? 0 = no 1 = yes	**Stroke:** you had a stroke. 0 = no 1 = yes
Heart disease: coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes	**PhysActivity**: physical activity in past 30 days - not including job 0 = no 1 = yes	**Smoker**: Have you smoked at least 100 cigarettes in your entire life, yes or no? 0 = no 1 = yes	**Veggies:** Consume vegetables 1 or more times per day 0 = no 1 = yes	**CholCheck:** 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
diabetes: 0 = no diabetes 1 = prediabetes 2 = diabetes (deleted prediabetes, and set diabetes as 1)	**HvyAlcoholConsumption**: Heavy drinkers (adult men < 14 drinks/week and women > 7 drinks/week) 0 = no 1 = yes	**Income:** Income scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more	**Education:** Education level scale 1-6 1 = Never attended school or only kindergarten 2 = elementary etc.	**NoDocbcCost:** Was there a time you didn't go to the doctor because of cost? 0 = no 1 = yes

Data visualization:

We first remove the pre-diabetes response which represents '2' and reset the diabetes response as '1' because "prediabetes is a serious health condition where blood sugar levels are higher than normal, but not high enough yet to be diagnosed as type 2 diabetes". The count plot are show below:





Before we get started on manipulating the data, we wanted to see whether or not the data was balanced overall. We used the count method and then made a histogram to display the dataset and as we can see above, the dataset was not balanced at all. This would be the equivalent of assigning the machine a task by telling it what to avoid rather than what to find. We might want a balanced data set because we want to give each result an equal amount of priority. For example, in our data set if the model guessed 100% of the points given had no diabetes, then we would have an on-paper accuracy of approx 85%. This seems to be okay, but our model would then be extremely weak at predicting if someone actually had diabetes.

In order to have the machine accurately differentiate between the two outcomes and put more stress on the actual features of the data rather than the balance of the data itself, we can do ‘upsampling’ to balance the response outputs (0 and 1). In the process of upsampling, observations from the minority class are randomly duplicated in order to amplify their signal. In this, we had a diabetes binary which would be 0 if the person does not have diabetes, and 1 if the person did have diabetes. The rest of the dataset was the exact same in terms of variables. This dataset is now balanced in terms of the number of people recorded as a 0 or 1 in the diabetes binary.

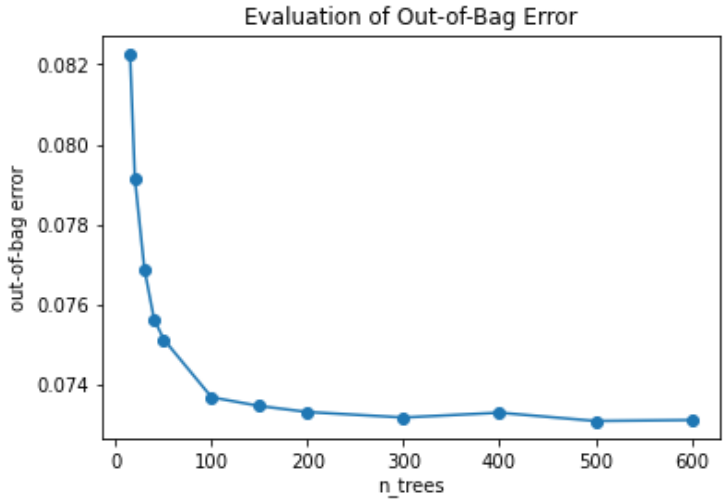
Some other changes we made to the dataset were removing the ‘MentHlth’ and ‘PhysHlth’ variables. We saw that these variables had a number of outliers and that they were not influencing our model as much. Removing them, in this case, would only help us avoid overfitting the model to the data. Additionally, we took the log(BMI) value for our BMI to help normalize the values. Because BMI turned out to be the most important variable in our model, we needed to normalize the outliers in order to enhance the learning accuracy of the model.

Methodology

Random Forest

We used the Random Forest method to train our model and predict whether a person has diabetes. We split these data into training and testing datasets. 70% is training data and 30% is testing data.

A Random Forest algorithm is a variety of the decision tree. It uses a multitude of constructed decision trees that predict the outcome of the response variable, and then chooses the most accurate ones to train the model on.



One big shortcoming of the Random Forest algorithm is the runtime. Because of the large number of decision trees, run before choosing appropriate paths to train the model, our algorithm runs for a much longer time compared to other algorithms. To reduce computing time, we first check the out-of-bag error. Out of bag error is the amount of error we need to account for in our model correlating to the number of trees in our forest. A variety of tree numbers were employed to fit the random forest model, and the out-of-bag error for each of the tree numbers was assessed. The error-tree graph is shown below. We chose to keep 300 trees for our data as a balance between training error and runtime of the algorithm.

From here, we decided to run the Random Forest model on our train data to create a train model. In this, we were able to see the following values of relative importance in the data: (Important Features table:)

Random Forest - Unbalanced Data		Random Forest - Balanced Data	
BMI	0.184041	BMI	0.205815
Age	0.121174	Age	0.148702
Income	0.097223	GenHlth	0.115439
PhysHlth	0.082688	Income	0.100602
GenHlth	0.073130	HighBP	0.081269
Education	0.068828	Education	0.069241
MentHlth	0.062810	HighChol	0.042667
HighBP	0.044329	Fruits	0.031469
Smoker	0.032852	Smoker	0.031407
Fruits	0.032796	DiffWalk	0.029333
HighChol	0.028009	Sex	0.026114
Sex	0.028005	PhysActivity	0.025161
Veggies	0.026243	Veggies	0.024424
PhysActivity	0.025945	HeartDiseaseorAttack	0.018851
DiffWalk	0.025143	NoDocbcCost	0.014238
HeartDiseaseorAttack	0.019681	Stroke	0.010887
NoDocbcCost	0.014851	HvyAlcoholConsump	0.009720
Stroke	0.012359	AnyHealthcare	0.008534
AnyHealthcare	0.008349	CholCheck	0.006125
HvyAlcoholConsump	0.007942		
CholCheck	0.003601		

In a random forest algorithm, like a decision tree, some variables have a higher relative impact on the final decision than others. The above shows us precisely the same, where higher relative influence means a spot closer to the tree root and a lower

influence is closer to the tree leaves showing the model's decision. Working on the unbalanced data also gave us the same order for the variables, albeit different relative importance values for each variable. From here, we ran the models on the testing data to check for their respective accuracies in making the decision on whether someone has diabetes. These results are shown below in the results section for part 2.

Logistic regression

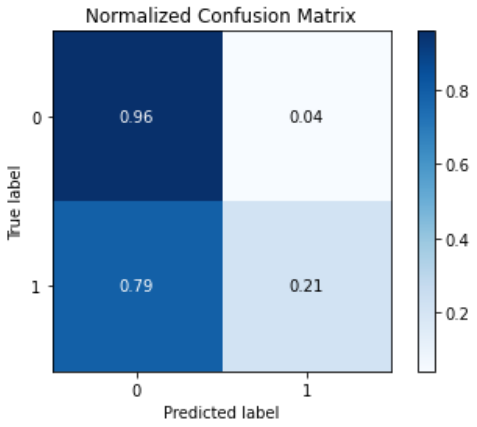
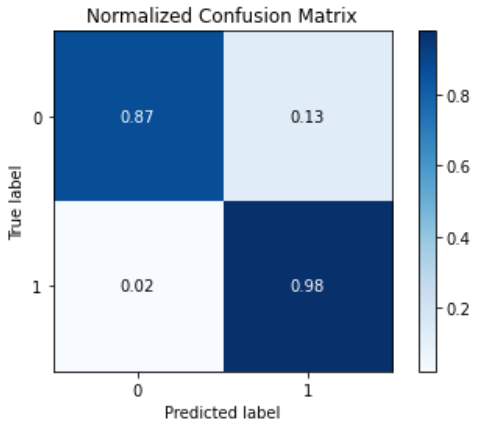
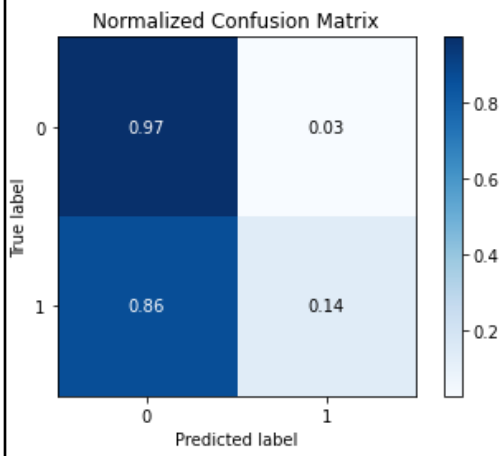
	coef	std err	z	P>|z|	[0.025	0.975]
HighBP	0.9028	0.017	52.359	0.000	0.869	0.937
HighChol	0.6289	0.016	39.402	0.000	0.598	0.660
CholCheck	-1.0175	0.034	-30.189	0.000	-1.084	-0.951
BMI	0.0216	0.001	22.577	0.000	0.020	0.024
Smoker	-0.1799	0.015	-11.717	0.000	-0.210	-0.150
Stroke	0.1242	0.030	4.086	0.000	0.065	0.184
HeartDiseaseorAttack	0.4144	0.021	19.431	0.000	0.373	0.456
PhysActivity	-0.2644	0.016	-16.222	0.000	-0.296	-0.232
Fruits	-0.0935	0.016	-5.893	0.000	-0.125	-0.062
Veggies	-0.1981	0.018	-11.046	0.000	-0.233	-0.163
HvyAlcoholConsump	-0.9171	0.046	-20.006	0.000	-1.007	-0.827
AnyHealthcare	-0.6774	0.030	-22.223	0.000	-0.737	-0.618
NoDocbcCost	-0.4081	0.027	-15.208	0.000	-0.461	-0.356
GenHlth	0.2537	0.008	31.336	0.000	0.238	0.270
MentHlth	-0.0088	0.001	-8.901	0.000	-0.011	-0.007
DiffWalk	0.3008	0.019	15.532	0.000	0.263	0.339
Sex	0.1359	0.016	8.718	0.000	0.105	0.166
Age	0.0212	0.003	7.447	0.000	0.016	0.027
Education	-0.3207	0.007	-44.318	0.000	-0.335	-0.307
Income	-0.0931	0.004	-23.208	0.000	-0.101	-0.085

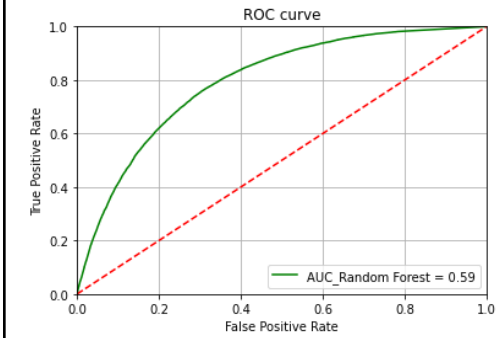
Based on our model summary, Income is a significant factor that influences whether a person has diabetes. The estimated log-odds ratio of getting diabetes, β_{Income} is -0.0931 between low income and high income. (Income scale 1 = less than $10,000 , 5 = less than $35,000, 8 = $75,000 or more) That means higher income people have a lower probability to get diabetes compared with low income people.

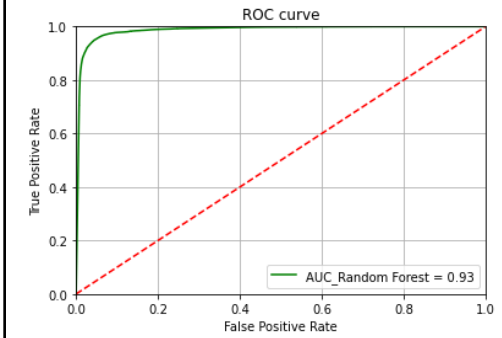
To make sure that we use the best model in terms of prediction both through intuition and accuracy, we decided to also train a separate logistic regression model on the same training data.

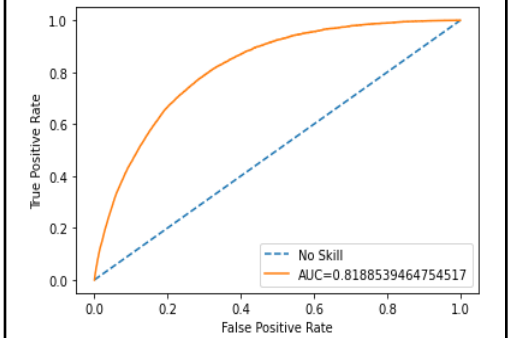
Now, we use logistic regression to predict if the person has diabetes or not. Above, we can see the relative weights for the regression model assigned to each variable in order to help make the prediction.

Results

Random Forest - Unbalanced Data	Random Forest - Balanced Data	Logistic Regression
		







Looking at the table above displaying our results we can make the following conclusions for Part II:

First, in terms of the confusion matrix, we can see a stark difference in the Balanced Random Forest in comparison to the Logistic and Unbalanced Random Forest models.

Looking more closely at the Unbalanced Random Forest and the Logistic models, we see that both predicted the probability of not having diabetes very accurately. The problem comes in predicting if someone does have diabetes. In these models, we have a high value for Type II error, meaning a false negative report on diabetes. This is very harmful in terms of our dataset as untreated diabetes can lead to gangrene and death. We see the same translate into our AUC curves in the bottom graphs. Both curves are showing a lower amount of resilience to Type I and Type II errors.

Now looking at the Balanced Random Forest, we see that the Type I and Type II errors are significantly lower than the error rates in the other models, as well as in the model itself compared to the number of accurately guessed results. We have a near perfect detection of diabetes in the model, as well as a very close to accurate amount of people who are accurately detected as not having diabetes. Here, we noticed a slight increase in the Type I error slightly higher in the Balanced Model rather than the unbalanced model. In this case, however, a Type I error is not a major concern, and in fact may be helpful in diabetes prevention in terms of this study. A Type I error in this case would be defined as being diagnosed with diabetes when the subject does not have diabetes. In this case, if the subject is discerned of having diabetes, they can follow steps to be careful about blood sugar, and therefore take preventative care for diabetes. We see the same in the AUC graph, where the model is near completely resilient to Type I and Type II errors.

Part 3 - What is the relation between price and a product's ingredients?

After using the random forest in part two, we figure out the important features that affect diabetes, we are interested in the parameter of 'income' (ranked #3) to see if it affects the probability of an accurate diabetes diagnosis. Even if we are aware that price cannot accurately represent "income," it can nonetheless show people some pictures of it since we all know low-income groups cannot always afford expensive goods. Thus, we are still interested in how nutrition affects the price of grocery store items. We are expecting a trend where the higher the unhealthy ingredients in food, the lower the price will be.

Data collection & cleaning

To achieve this goal, we use the web-scraping method to get the data from a local grocery store: Safeway. We chose bread as our main research target. Because bread is typically consumed as a key source of food in the US and comes in a wide range of varieties. Also, based on the personal experiences of our group members, we assumed that the brand of bread has little impact on its pricing.

Data Description

Price - our response variable
All predicting variables are in Amount Per Serving size : **Calories** - in cal, **Total Fat**- in g, **Saturated Fat** - in g, **Cholesterol** - in mg, **Sodium** - in mg, **Potassium** - in mg, **Total Carbohydrate** - in g, **Dietary Fiber** - in g, **Total Sugars** - in g, **Added Sugars** - in g, **Protein** - in g, **Calcium** - in mg, **Iron** - in mg

Methodology

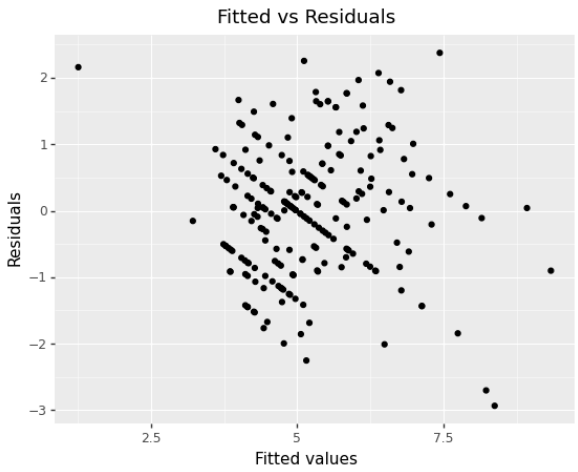
We chose a linear regression model to identify the most statistically significant variables and drop those that are not significant. We first try to fit all current variables, mentioned above, into the linear regression model to see if this model is significant and meets all our needs.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.361
Model:	OLS	**Adj. R-squared:**	0.322
Method:	Least Squares	**F-statistic:**	9.201
Date:	Mon, 05 Dec 2022	**Prob (F-statistic):**	5.77e-15
Time:	18:27:05	**Log-Likelihood:**	-404.35
No. Observations:	226	**AIC:**	836.7
Df Residuals:	212	**BIC:**	884.6
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.3477	0.279	19.181	0.000	4.798	5.897
Total Fat	0.0143	0.011	1.267	0.206	-0.008	0.037
Cholesterol	0.0901	0.021	4.290	0.000	0.049	0.131
Sodium	-0.0004	0.002	-0.239	0.811	-0.003	0.003
Potassium	0.0024	0.003	0.813	0.417	-0.003	0.008
Total Sugars	0.2602	0.081	3.208	0.002	0.100	0.420
Dietary Fiber	0.3325	0.055	6.085	0.000	0.225	0.440
Calcium	-0.0099	0.002	-4.236	0.000	-0.014	-0.005
Iron	-0.0034	0.010	-0.343	0.732	-0.023	0.016
Calories	-0.0104	0.006	-1.721	0.087	-0.022	0.002
Added Sugars	-0.1232	0.088	-1.404	0.162	-0.296	0.050
Total Carbohydrate	0.0315	0.026	1.219	0.224	-0.019	0.082
Saturated Fat	-0.0112	0.023	-0.485	0.628	-0.057	0.034
Protein	-0.1467	0.071	-2.073	0.039	-0.286	-0.007

Omnibus:	0.050	**Durbin-Watson:**	1.913
Prob(Omnibus):	0.975	**Jarque-Bera (JB):**	0.060
Skew:	-0.033	**Prob(JB):**	0.970
Kurtosis:	2.955	**Cond. No.**	838.



From this initial model, two plots are generated in order to check if our data meet the assumptions for the linear models. We use the fitted vs residuals plot to check if it fits normality and qq-plot for linearity.

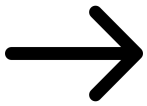
The fitted vs residuals plot above shows us that our residuals are evenly distributed above and below the zero line; homoscedasticity is not shown in this plot. The QQ-Plot presents all our data dots sticking to the red linear line, so we know our residuals are normally distributed.

The next assumption we check is if the variables are independent of each other. A heatmap of all the variables in a correlation matrix is shown below.

	price	Total Fat	Cholesterol	Sodium	Potassium	Total Sugars	Dietary Fiber	Calcium	Iron	Calories	Added Sugars	Total Carbohydrate	Saturated Fat	Protein
price	1.000000	0.092130	0.327472	-0.056050	0.187481	0.281560	0.265398	-0.201806	-0.095684	-0.003430	0.208652	-0.035252	0.175624	-0.029637
Total Fat	0.092130	1.000000	0.077630	0.177951	0.250795	0.139890	0.145904	0.138614	0.158452	0.204871	0.228090	0.134382	0.095476	0.102004
Cholesterol	0.327472	0.077630	1.000000	-0.078999	0.212002	0.292152	-0.067242	-0.031017	-0.049390	0.112459	0.256214	-0.076486	0.608790	-0.030728
Sodium	-0.056050	0.177951	-0.078999	1.000000	0.299980	0.243857	0.347280	0.094541	0.226288	0.748443	0.208559	0.700250	0.010648	0.615598
Potassium	0.187481	0.250795	0.212002	0.299980	1.000000	0.417768	0.393274	0.182005	0.138291	0.516971	0.334394	0.406710	0.174063	0.482576
Total Sugars	0.281560	0.139890	0.292152	0.243857	0.417768	1.000000	0.083173	-0.021657	0.070094	0.546177	0.849842	0.535998	0.353663	0.224609
Dietary Fiber	0.265398	0.145904	-0.067242	0.347280	0.393274	0.083173	1.000000	0.203787	0.104896	0.225807	0.081807	0.132566	0.042332	0.503930
Calcium	-0.201806	0.138614	-0.031017	0.094541	0.182005	-0.021657	0.203787	1.000000	0.167471	0.067637	0.023922	0.059437	0.014672	0.152557
Iron	-0.095684	0.158452	-0.049390	0.226288	0.138291	0.070094	0.104896	0.167471	1.000000	0.212608	0.141135	0.173368	0.028963	0.250200
Calories	-0.003430	0.204871	0.112459	0.748443	0.516971	0.546177	0.225807	0.067637	0.212608	1.000000	0.483221	0.897486	0.280721	0.629323
Added Sugars	0.208652	0.228090	0.256214	0.208559	0.334394	0.849842	0.081807	0.023922	0.141135	0.483221	1.000000	0.465776	0.377737	0.086550
Total Carbohydrate	-0.035252	0.134382	-0.076486	0.700250	0.406710	0.535998	0.132566	0.059437	0.173368	0.897486	0.465776	1.000000	0.073245	0.523941
Saturated Fat	0.175624	0.095476	0.608790	0.010648	0.174063	0.353663	0.042332	0.014672	0.028963	0.280721	0.377737	0.073245	1.000000	0.075380
Protein	-0.029637	0.102004	-0.030728	0.615598	0.482576	0.224609	0.503930	0.152557	0.250200	0.629323	0.086550	0.523941	0.075380	1.000000

We will take those variables that are highly correlated, in this case taking the variables that are more red in the heatmap and compute their variance inflation factor(VIF) to find multicollinearity. These variables are Calories, Total Carbohydrate, Total Sugars, Added Sugars, Sodium, Protein, Potassium, and Saturated Fat. We used the threshold of >5 when looking at the VIFs. We dropped the variable “Calories” since its VIF is above this threshold and we recomputed VIFs afterward to see if there is still any variable’s VIF above 5.

	Variable	VIF
0	Calories	10.714543
4	Total Carbohydrate	7.119850
5	Total Sugars	4.313632
3	Added Sugars	4.035020
1	Sodium	2.930893
7	Protein	2.226553
2	Potassium	1.637669
6	Saturated Fat	1.628396



	Variable	VIF
4	Total Sugars	4.312239
2	Added Sugars	4.018837
3	Total Carbohydrate	2.858007
0	Sodium	2.499207
6	Protein	2.099510
1	Potassium	1.518412
5	Saturated Fat	1.220264

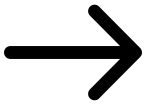
Now that the variables have VIFs < 5, we refit the linear regression

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.352
Model:	OLS	**Adj. R-squared:**	0.315
Method:	Least Squares	**F-statistic:**	9.632
Date:	Mon, 05 Dec 2022	**Prob (F-statistic):**	6.86e-15
Time:	18:27:07	**Log-Likelihood:**	-405.92
No. Observations:	226	**AIC:**	837.8
Df Residuals:	213	**BIC:**	882.3
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.2747	0.277	19.054	0.000	4.729	5.820
Total Fat	0.0127	0.011	1.126	0.261	-0.010	0.035
Cholesterol	0.0859	0.021	4.099	0.000	0.045	0.127
Sodium	-0.0014	0.001	-0.963	0.336	-0.004	0.001
Potassium	0.0011	0.003	0.383	0.702	-0.004	0.007
Total Sugars	0.2670	0.081	3.282	0.001	0.107	0.427
Dietary Fiber	0.3429	0.055	6.285	0.000	0.235	0.450
Calcium	-0.0095	0.002	-4.081	0.000	-0.014	-0.005
Iron	-0.0037	0.010	-0.375	0.708	-0.023	0.016
Added Sugars	-0.1352	0.088	-1.538	0.125	-0.308	0.038
Total Carbohydrate	-0.0019	0.017	-0.111	0.912	-0.036	0.032
Saturated Fat	-0.0265	0.021	-1.240	0.217	-0.069	0.016
Protein	-0.1806	0.068	-2.645	0.009	-0.315	-0.046

Omnibus:	0.144	**Durbin-Watson:**	1.918
Prob(Omnibus):	0.931	**Jarque-Bera (JB):**	0.024
Skew:	0.009	**Prob(JB):**	0.988
Kurtosis:	3.047	**Cond. No.**	740.



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.326
Model:	OLS	**Adj. R-squared:**	0.311
Method:	Least Squares	**F-statistic:**	21.29
Date:	Mon, 05 Dec 2022	**Prob (F-statistic):**	2.36e-17
Time:	18:27:07	**Log-Likelihood:**	-410.31
No. Observations:	226	**AIC:**	832.6
Df Residuals:	220	**BIC:**	853.1
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.2081	0.233	22.332	0.000	4.749	5.668
Cholesterol	0.0773	0.016	4.790	0.000	0.046	0.109
Total Sugars	0.1463	0.041	3.607	0.000	0.066	0.226
Dietary Fiber	0.3384	0.049	6.850	0.000	0.241	0.436
Calcium	-0.0097	0.002	-4.230	0.000	-0.014	-0.005
Protein	-0.2000	0.051	-3.901	0.000	-0.301	-0.099

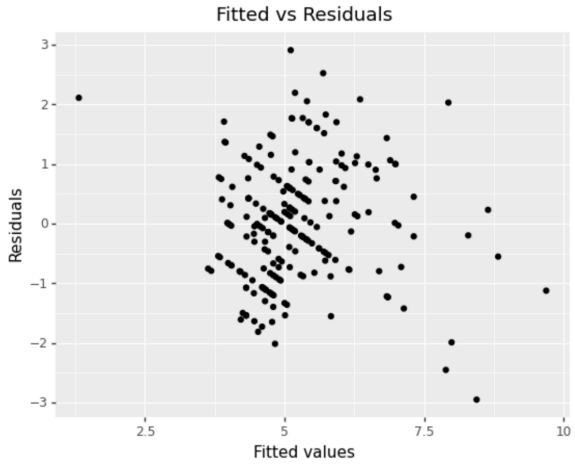
Omnibus:	0.792	**Durbin-Watson:**	1.878
Prob(Omnibus):	0.673	**Jarque-Bera (JB):**	0.777
Skew:	0.141	**Prob(JB):**	0.678
Kurtosis:	2.946	**Cond. No.**	136.

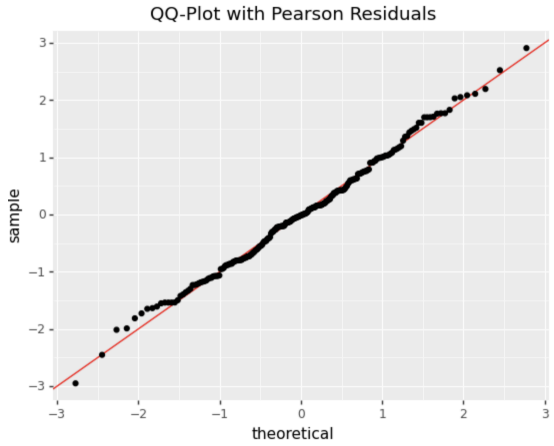
In this model, we dropped highly correlated variables. However, we still see many variables with large p-values, which means they are insignificant. In order to drop these insignificant variables, we use BIC backward selection and choose a model with the lowest BIC. This current model has a BIC of 882.311

After we compare every possible BICs, the one with the lowest BIC is 853.146, thus we will choose this model since it appears to be the most optimal. This model includes cholesterol, total sugars, dietary fiber, calcium, and protein.

Now we get our **final linear regression model**

This final model is now the most optimal since all variables have a p-value less than 0.05. Then we use fitted vs residuals plot and qq-plot to check for homoscedasticity and normality again.





The fitted vs residuals plot shows us once again that the residuals are evenly distributed with no patterns emerging. The qq-plot shows us a linear trend in the plot. These two plots indicate homoscedasticity and normality of our data are fine.

We are expecting to see a positive trend between good nutrients(e.g.Dietary Fiber) and price, while a negative trend is between bad nutrients(e.g.Saturated Fat, Added Sugars) and price. However, there are numerous nutrients that are excluded from our final model. Therefore, our conclusion is only limited to variables that exist in our final model. Cholesterol, Total Sugars, and Dietary Fiber have a positive linear relationship with price. While Calcium and Protein have a negative linear relationship with price.
To sum up, it is hard for us to conclude if there is any linear relationship between good nutrition or bad nutrition and price. Based on our model, we only can make conclusions for each nutrient in our model, and they cannot represent all the good or bad nutrients.

Overarching Results

For this project, we created two models. One for the best indicators of diabetes and predicting diabetes outcomes and the second for analyzing the relationship between price and nutritional content of bread. Part one involved using API to get data from the CDC website. Part three had us scraping data from the Safeway website in order to obtain prices and product names.

For our first model using logistic regression and random forest, we found the top five indicators for predicting type II diabetes (+ to indicate positive association and - for negative association) to be high blood pressure (+), age (+), general health assessment (+), their body mass index (+), and income (-). When put to the test, the ROC curve for our logistic regression gave us an area of 0.81. Our random forest algorithm had an ROC curve area of 0.93. Random forest performed substantially better than our logistic regression model.

In part three, we created a model that used linear regression to find a relationship between price and nutritional content. In our model, the five best indicators were cholesterol (+), total sugars (+), dietary fiber (+), calcium (-), and protein (-). It is hard for us to make a final determination on the nutritional health of these bread and their price due to many of our variables not being utilized in the final model.

Discussion

In part two we created a model that predicted if a patient had type II diabetes or not. Our final model using random forest outperformed our logistic regression model. However, both had decent predictive power. Random forest was more difficult to implement. At first, our data was unbalanced and this resulted in our first attempt at the random forest to produce a very weak model.

The results from our models in part two were not surprising and showed a similar pattern to what was shown in part one. Notably, income and age both were two of the most influential predictive variables. The notion that as income decreases and age increases the higher chance a person will have diabetes. As well as that, the other top indicators, BMI, high blood pressure, and a general health assessment, are not terribly surprising results. Specifically for BMI and high blood pressure, both of these are well-known causes of a multitude of diseases. A self-assessment of general health is a bit more interesting of a result because bias can skew the responses of participants.

In part three, our prediction was that the price of bread would rise as good nutritional content increased. It is hard to determine this relationship after our analysis due to many of the predictor variables being deemed not as useful during the selection process. Also, scraping data was the major process of this part. We tried many grocery store websites and finally decided to choose

Safeway data since that was the only one we successfully scraped. Maybe we should also try the API method, which might bring us a different outcome.

Although we were able to find the most optimal model for linear regression, some of our results don't agree with our prediction. For instance, the total sugars and cholesterol in bread have a positive relationship with price. We would expect that these generally bad indicators of nutrition would make the price of bread go down, but that was not the case. A reason for this could be that there was an inclusion of dessert-like items within our analysis. These special bread items generally cost more and could skew our results.

For our good indicators for health, our model did show that dietary fiber did show a positive relationship with the price; this falls in line with our prediction. Protein and calcium, however, showed a negative relationship which does not fall in line with our prediction.

Overall, with this model in part three, we can see certain variables have a positive or negative relationship with price, but we don't have enough information to establish a relationship between healthy nutrition and price. It is possible that in a future investigation, using a different product and its prices that are better analyzed would yield clearer results.

Appendix: **code uploaded to github**