ML Based Predictive Model for Chip Sizes Using Raw Potato Size Data

Pallavi Gupta

University of Missouri(Columbia)

R&D - Data Science and Analytics

Date - August 6, 2020

About Me

Background

- Born and raised in Shimla, India (a town in Himalayas)
- Family in India
- Lays Chips are my favorite PepsiCo product.

Education

- Currently pursuing PhD in Informatics from University of Missouri(Mizzou)
- M.S. (Bioinformatics & Computational Biology) from Saint Louis University(SLU)
- Bachelors in Engineering (Biotech) from India

Experience

- Currently working as a Graduate Research fellow (Mizzou),since past 1 year..
- Worked as a Graduate Research assistant (SLU) for 2 years.
- ~5 years of work experience as a Software analyst in India.

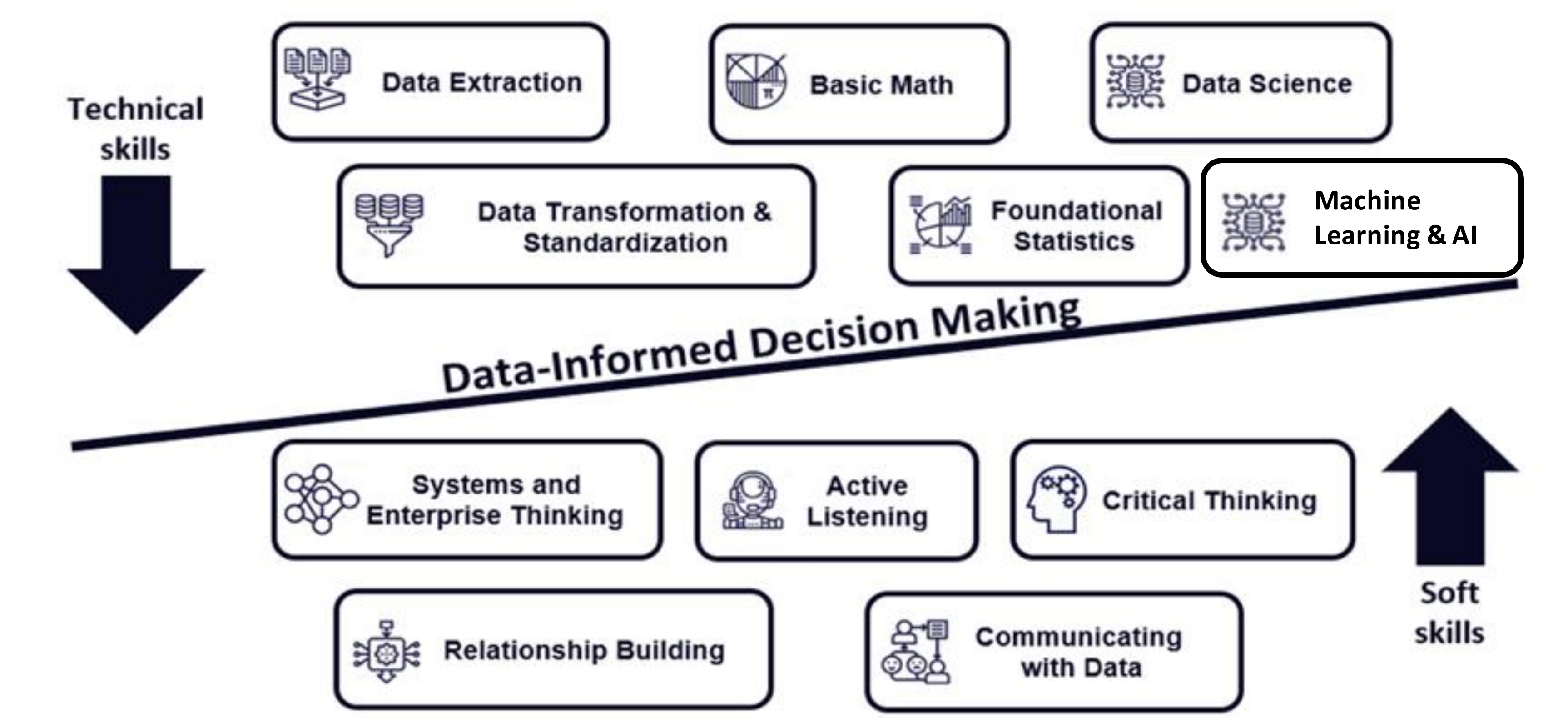
Interests

- Travelling and exploring new parts of the world.
- Love to explore new food(foodie)
- DIY creative decorations & cooking, when at home

Agenda

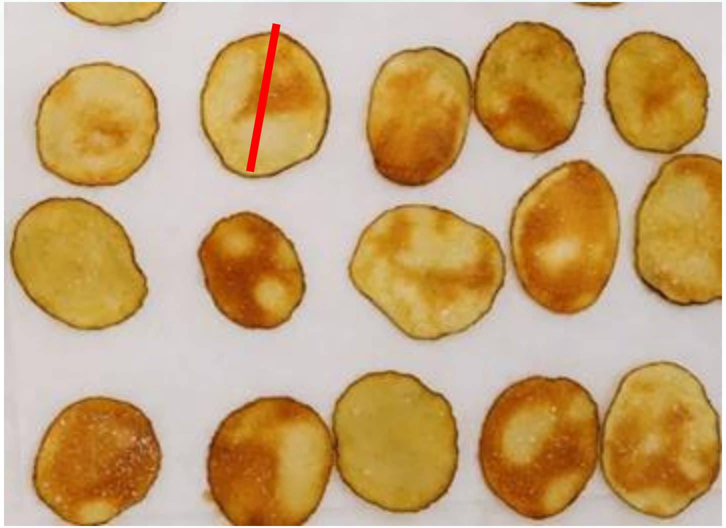
- Introduction
- Objectives
- Project Significance
- Methodology
- Results & Analysis
- Recommendations

Introduction - Data Science & Analytics



Objective

To build a machine learning based prediction model to be able to predict the %age of chips produced (having length >2.5 inches), from each truckload of potatoes.



Project Significance

- This project would help the chips production team to estimate the right quantity of small packets to order and avoid wastage.
- Small packets can fit small chips with L<2.5 inches.
- Hence 2.5 inch is our threshold.

Note: **length > 2.5inch - Large chips**

length < 2.5inch - Small chips

Methodology - ML Based approach

- **Step1:** Data exploration - Get to know the data
- **Step2:** Data exploration - Biases in data that can affect results
- **Step3:** Possible solutions - Based on data explorations
- **Step4:** Feature engineering and data structurization
- **Step5:** Train and Test models (with binarization approach)
- **Step6:** Train and Test models (with multi-class approach)

Methodology - ML Based approach

- **Step1:** Data exploration - Get to know the data
- **Step2:** Data exploration - Biases observed in data
- **Step3:** Possible solutions - Based on data explorations
- **Step4:** Feature engineering and data structurization
- **Step5:** Train and Test models with binarization model
- **Step6:** Train and Test models with multiclass model

Imagining potato population in a truckload

Assumptions in each truckload

- Each truckload will have all the varieties/types/sizes of potato.
- Each truckload will produce multiple sizes of chips including small chips to large chips.
- Considering the consensus of population of potatoes in a truckload we can predict probabilities or %ages of chips(small or large) that can be produced more in number.
- We have got data from ~700 truckloads from 3 months.
- Each truckload sums up into 1 data point in the data we have.

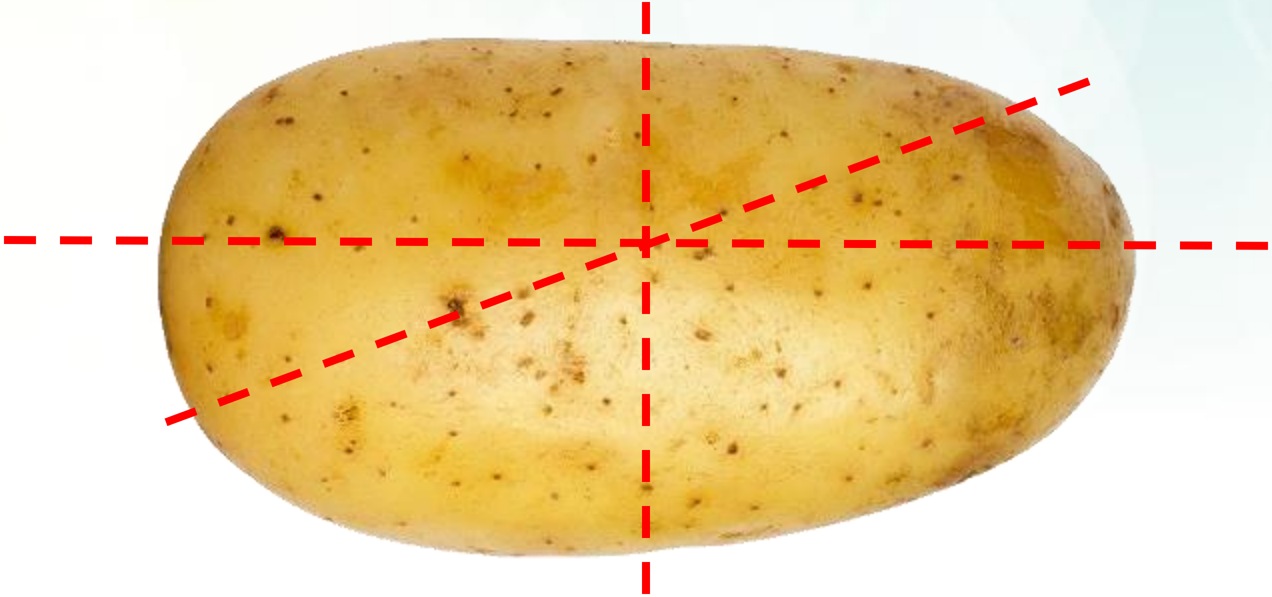
Methodology - ML Based approach

- **Step1:** Data exploration - Get to know the data
- **Step2:** Data exploration - Biases in data
- **Step3:** Possible solutions - Based on data explorations
- **Step4:** Feature engineering and data structurization
- **Step5:** Train and Test models with binarization model
- **Step6:** Train and Test models with multiclass model

Methodology

Step2: Data exploration - Biases in data

Bias1: Angle of slicing



Types of population of potatoes vs angle bias

- **Type1 Population** : [L and D] <2.5 in (Small Potatoes)
    - Produces only small chips (having L< 2.5 inches)
- **Type2 Population** [L and D] >2.5in (Bulky Potatoes)
    - Produces maximum large chips (having L>2.5 inches)
- **Type3 Population** : [L >2.5 and D<2.5] in **(TUBERS)**
    - Most affected by angle bias
    - If it is sliced along the length, large chips (L >2.5 in) will be produced
    - If it is sliced along the diameter, small chips (L <2.5 in) will be produced

Bias2: Kind of chips produced from each workload would affect number of chips

No information on chips type is provided.

- **Bias3: Dimensions are from Image data (2D)**
  - We have length, diameter & area for potatoes and length & area for chips
  - It does not consider factors like curves in chips, tapering ends of potatoes, maximum or minimum Diameter in potato, etc.
  - Hence, data is incomplete.
- **Bias 4: Lesser data points**
  - Currently we just have 3 months data summing to ~700 data points.
- **Bias 5: Missing image data**
  - Among 700 data points we have some missing data as well

- **Bias 5: Population to population data**
    - Current data considers that a sample population of potato with varying sizes would produce a sample population of chips with varying sizes.
    - However, given a potato with certain dimensions, how many corresponding chips would be produced? Not fully known..
    - **1 truckload ultimately sums up into 1 datapoint**

Methodology - ML Based approach

- **Step1:** Data exploration - Get to know the data
- **Step2:** Data exploration - Biases in data
- **Step3: Possible solutions - Based on data explorations**
- **Step4:** Feature engineering and data structurization
- **Step5:** Train and Test models with binarization model
- **Step6:** Train and Test models with multiclass model

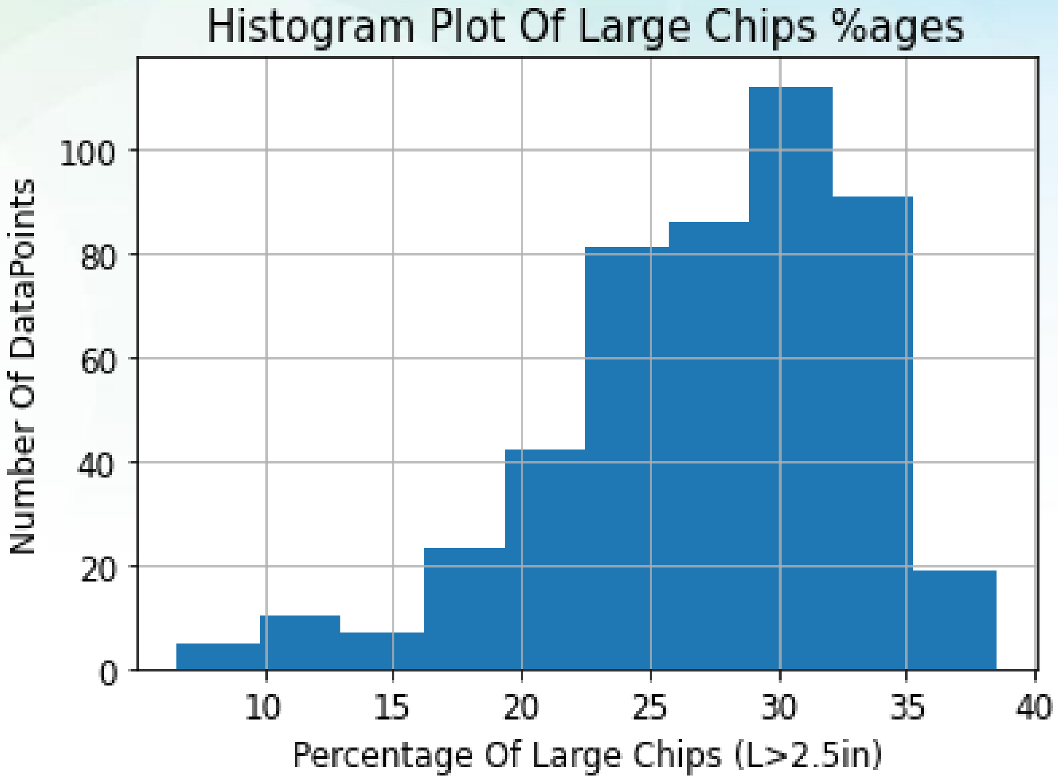
Methodology

Step3: Possible solutions - Based on data explorations

- Prediction of exact %age of large chips (having length >2.5 in) is practically challenging due to the biases.
- Possible models/approaches can be classification based:
  - Approach 1 **Binary Classification prediction model**
  - Approach 2 **Multi classification prediction model**
- <u>Feature engineering and data structurization is must</u> before modeling the data.

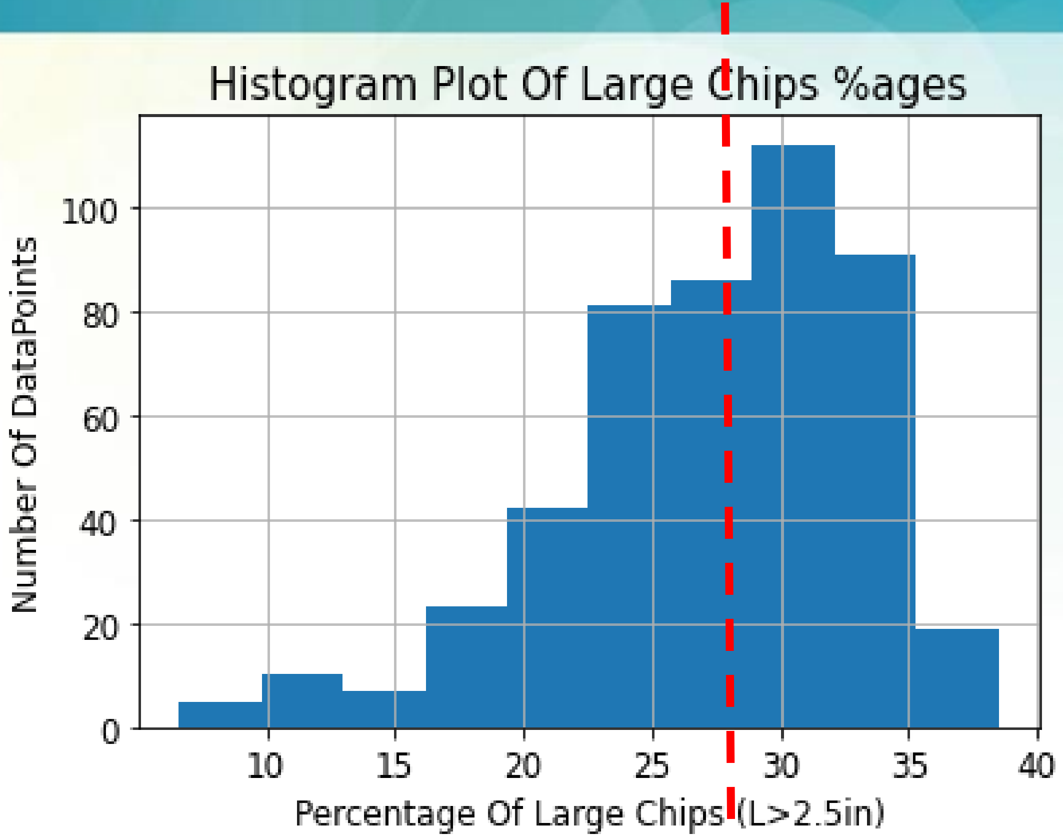
Few more take away from data:

- After data cleaning, only 476/700 data points left.
- Out of 476 data points, maximum %age of large chips (L>2.5in) obtained from any truck load was <40%
- None of the truckload could produce >40% of large chips



Approach1: Binary Classification prediction model (Broader Classification)

- Binarize the data points into categories of %age of large chips (L > 2.5in)
  - 0-28% as category 1
  - >=28% as category 0
- With this method data looked balanced in two categories with 28 as cutoff for binarizing categories. Out of total data Points (476)
  - Number of datapoints in >=28% slab(0)==> 247
  - Number of datapoints in <28% slab(1)==> 229



Histogram Plot Of Large Chips %ages
Number Of DataPoints
0
20
40
60
80
100
10
15
20
25
30
35
40
Percentage Of Large Chips (L>2.5in)

Approach2: Multi classification prediction model (Narrowed Classification)

- Define categories for % of large chips (L>2.5 in), and then make predictions falling in either of these slabs.
    - 0-10%chips defined as category 1
    - 10- 20% as category 2,
    - 20-30% as category 3,
    - 30-40% as category 4
    - As per current data, there are 0 data points >40%.

- Data is imbalanced when categorized in multi-classes.
- Out of 476 data points available for training the model
  - 6 data points in 0-10% slab of chips having L >2.5inch
  - 46 data points in 10-20% slab
  - 229 data points in 20-30% slab
  - 195 data points in 30-40% slab
- Currently, data is insufficient for this approach, still tried.

Methodology - ML Based approach

- **Step1:** Data exploration - Get to know the data
- **Step2:** Data exploration - Biases in data
- **Step3:** Possible solutions - Based on data explorations
- **Step4:** Feature engineering and data structurization
- **Step5:** Train and Test models (with binarization approach)
- **Step6:** Train and Test models (with multi-class approach)

Methodology

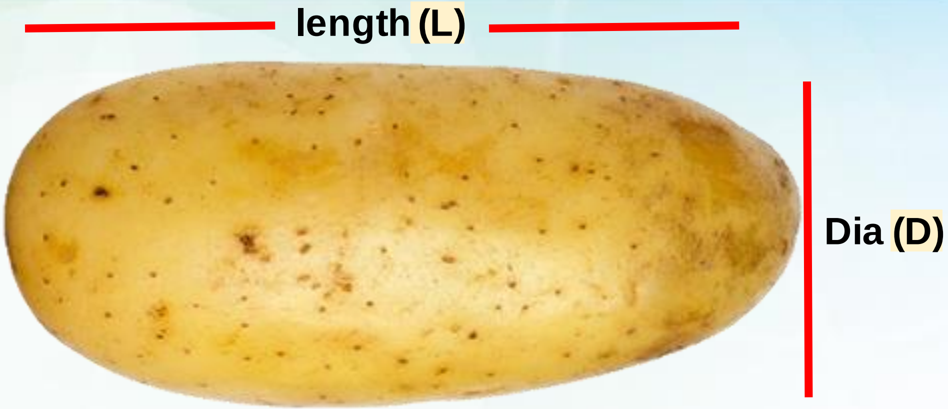
Step4: Feature engineering and data structurization

- Engineering the right features to train the model with the statistics of data is very important.
- Less features carrying more weightage is good to have.
- Having features that correlate in certain way helps to make predictions in the right direction.
- In current case, I engineered total 7 features.
- In those 7 features, I also tried to get over angle bias of slicing using L/D ratio for TUBERS category.

Methodology

Step4: Feature engineering and data structurization

- **Calculation of L/D ratio for tubers (L>=2.5in and D=<2.5in) (prone to max. Angle bias)**



- L/D ratio will help reduce the angle bias, as ML model can learn from the examples we provide.
- **Tubers having L/D ratio <1.25** will show very less angle bias. (If cut diametrically or longitudinally)(typically roundish potatoes)
- **Tubers having L/D >1.25** are expected to show major angle bias.

Methodology

Set of 7 features created

- Feature1: %age of small potatoes (L AND D) <2.5 in
- Feature2: %age of bulky potatoes (L AND D) >2.5 in
- Feature3: %age of TUBER potatoes (L >=2.5 AND D=<2.5in)
- Feature4: %age of TUBER potatoes having L/D ratio == 1 (round)
- Feature5: %age of TUBER potatoes having L/D ratio ~ [1-1.25] (roundish)
- Feature6: %age of TUBER potatoes having L/D ratio ~ [1.25-2.5] (Fairly long)
- Feature7: %age of TUBER potatoes having L/D ratio >2.5 (very long)
- Target Column: %age of large chips produced [With categorized %age]

Methodology - ML Based Approach

- **Step1:** Data exploration - Get to know the data
- **Step2:** Data exploration – Biases in data
- **Step3:** Possible solutions - Based on data explorations
- **Step4:** Feature engineering and data structurization
- **Step5:** Train and Test models (with binarization approach)
- **Step6:** Train and Test models (with multi-class approach)

Methodology

Step5: Train and Test models (with binarization approach)

Trained the following models and tested them for accuracy:

- **Logistic Regression:** With cross validation=10, C=1.
- **Neural Networks:** Sequential with following architecture

10 inputs (Input layer)	->	[8 hidden nodes] (Hidden layer)	->	binary output (Sigmoid layer)

- **Random Forest Classifier:** With grid search and cross validation=10. (Depth=5, n_estimator=100)
- **XGBoost Classifier:** With grid search and cross validation=10.

Methodology - ML Based Approach

- **Step1:** Data exploration - Get to know the data
- **Step2:** Data exploration - Biases in data
- **Step3:** Possible solutions - Based on data explorations
- **Step4:** Feature engineering and data structurization
- **Step5:** Train and Test models (with binarization approach)
- **Step6:** Train and Test models (with multi-class approach)

Methodology

Step6: Train and Test models (with multi-class approach)

Trained the following models and tested them for accuracy:

- **Logistic Regression:** With grid search and cross validation=10
  - C=0.1
- **Random Forest Classifier:** With grid search and cross validation=10
  - max_depth=10, n_estimators=50
- **XG Boost Classifier:** With grid search and cross validation=10
  - booster='gblinear'

Result and Analysis

	Binarization Approach		Multi-Class Approach	
	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
Logistic Regression	63.42%	67.70%	52.65%	56.30%
Random Forest Classifier	60.50%	62.50%	50.44%	48.74%
Sequential Neural Network	64.21%	66.67%	Not applicable due to less data in each category	
XGBoost Classifier	62.90%	67.71%	52.38	52.94%

Models

Result and Analysis

- Tried two classification approaches out of which binarization approach performed better.
- Multi-class approach suffered due to insufficient data.
- Models trained with both the approaches exhibited average accuracy for predictions.
- Models alone cannot make predictions betters. Data is also equally important.
- Hence, I have some recommendations going forward in future.

Recommendations

- **Insufficient data:** At least 6-9 months data required to train the model sufficiently, in current case.
- **Garbage-in garbage out:** Data Quality has to be improved. Reduce missing data.
- **Right Data:** Having some **experimental lab data examples** would be helpful
    - Like given some standard potato shapes and sizes, how many %age of large chips (L>2.5in) would be produced?
- When enough data is available, neural networks approach would be effective with multi-classification approach.

Acknowledgements

- **James Yuan**
  Sr. Director – Data Science & Analytics

- **Hua Xu**
  Senior Principal Scientist - Data Science & Analytics

- **Sonchai Lange**
  R&D Engineer - Process & Product Development

Internship Learnings

- **Success metric**
  First and foremost understand the success metric for building any predictive model.
- **Understand the data**
  Data might not be complete always. Hence, it's a duty of a data scientist to explore, talk to cross functional team and decide if more data is required.
- **Discussions are must**
  Collaborative interactions within own team as well as with cross functional teams are must to be productive and successful in the endeavor
- **Be open in thoughts**
  Be open and confident about own perspective. Also, be open to receive any kind of suggestions/critiques.
- **Along with projects, networking is also equally important.**

Questions?

APPENDIX

THANK YOU!!

PepsiCo R&D