# CS680 PROJECT REPORT
# FAST FOOD PREFERENCE PREDICTION FROM GROCERY ORDERS

**Shreyan Mahalanabis**
Student # 21204674
s2mahala@uwaterloo.ca

## ABSTRACT

This project aims to predict Canadian consumer preference towards fast-food using machine learning. The project proposal contains information about the motivation behind the project, related work in the research field, acquiring the data to train the model, prediction results and key takeaways. —-Total Pages: 6

## 1 MOTIVATION

The Canadian food services industry is a critical part of the economy, amassing over $95 billion CAD in operating revenue in 2023 (Statistics Canada., a). However, even though the industry has bounced back since the pandemic, the sector is still struggling with historically low profit margins (3.6%) (Statistics Canada., c). Multiple food service locations in metropolitan areas have shut down operations, but at the same time, certain businesses have thrived in the same environment. Further, with the large-scale adoption of third-party delivery apps, product sales have increased, and 11% of all sales are now through these apps compared to only 2% back in 2016 (Statistics Canada., b). The primary objective of this project is to utilize diverse datasets from Statistics Canada on consumer food preferences and grocery data from Instacart (Instacart) to identify consumer fast food preferences as an alternative to sensitive demographic data.

## 2 BACKGROUND AND RELATED WORK

The average Canadian household spending on restaurant food hit its peak in 2023 (Statistics Canada., a). According to an analysis conducted by the Public Health Agency of Canada with the Government's Foodbank initiative (Government of Canada.), a person's demographics played a role in how often they ordered fast food (Seale et al., 2023). Research on demographics affecting fast food intake has consistently shown that a person's age, ethnicity, and income are consistently able to predict their fast food consumption (Nimit N. Shah and Cheryl D. Fryar and Brian Tsai and Cynthia L. Ogden, 2025). According to UKCRC Centre for Diet and Activity Research (Burgoine et al., 2018), an individual in the low-income group was 1.54 times more likely to consume processed meats than someone in the higher income group (95% Confidence Interval). My project aims to use the Foodbank microdata and Instacart's grocery order data to analyze consumer grocery preferences in order to predict their fast food consumption and preference for fast-food cuisine without the use of demographics. Often, sensitive data of an individual consumer's income, age, etc, is not available. This research explores the viability of using grocery order data, which can be acquired through a grocery store's rewards program, as a predictor.

## 3 Data Acquisition and Processing

We use Foodbank microdata (Government of Canada.) from Health Canada to get information on what food an individual has eaten in the past week. This data was collected in 2014 and has 221 different food categories with questions on whether they had fast food and, if they did, what cuisine they had. The Instacart dataset (Instacart) has 33 million data points with the name of the grocery item, customer ID, and order ID. This data was published for the American market in the year 2017. There are 50 thousand grocery item names that are matched to the 221 categories using a hybrid approach of keyword matching and semantic matching using the language model all-mpnet-base-v2. This matches 72.2% of the dataset, and the rest are removed. This leaves us with 24 million data points. The Foodbank microdata has 32 thousand data points with 12% of participants eating fast food. We are trying to predict if they had fast food, and the fast food type (asian, indian, Burger, etc, or even multiple). After cleaning up datapoints where an individual mentioned eating fast food but refused to answer what they ate, we balance the dataset to avoid class imbalance, and we are left with 3,500 valid entries, with 30% of entries having fast food with an associated fast food type.

## 4 Methodology and Results (Grocery Data)

### 4.1 Methodology

A BERT-mini model is used for food category prediction using 204,583 purchase sequences across 195 categories. The dataset was split into training (173,895 samples) and validation (30,688 samples) sets. After removing 4 singleton classes, 182 categories were used for training.

#### 4.1.1 Model Architecture

The model consists of a BERT-mini backbone (11M parameters) with a custom classification head containing one hidden layer (512 units), batch normalization, and dropout (0.3).

#### 4.1.2 Training Configuration

AdamW optimizer is used with learning rate $5 \times 10^{-5}$, batch size 64, and trained for 15 epochs. The model was initialized with pre-trained weights and fine-tuned on the full dataset.

#### 4.1.3 Evaluation Metrics

Performance is evaluated using top-1, top-5, and top-10 accuracy metrics to account for the large label space (195 classes). We use the metric top-n accuracy to measure if the next item is in the top-n most probable outputs.

### 4.2 Results

#### 4.2.1 Overall Performance

The model achieved the best validation accuracy of 22.08% at epoch 14. Final metrics were: top-1 accuracy 21.99%, top-5 accuracy 49.70%, and top-10 accuracy 63.11%.

Table 1: Performance Metrics

| **Metric** | **Value** |
|---|---|
| Best Validation Accuracy | 22.08% |
| Final Top-1 Accuracy | 21.99% |
| Final Top-5 Accuracy | 49.70% |
| Final Top-10 Accuracy | 63.11% |
| Training Samples | 173,895 |
| Validation Samples | 30,688 |

### 4.2.2 TRAINING PROGRESS

The model showed consistent improvement over 15 epochs (Figure 1a,b). Validation accuracy increased from 20.22% to 22.08%, while training loss decreased by 9.6%. The learning rate schedule included warmup and linear decay (Figure 1d).

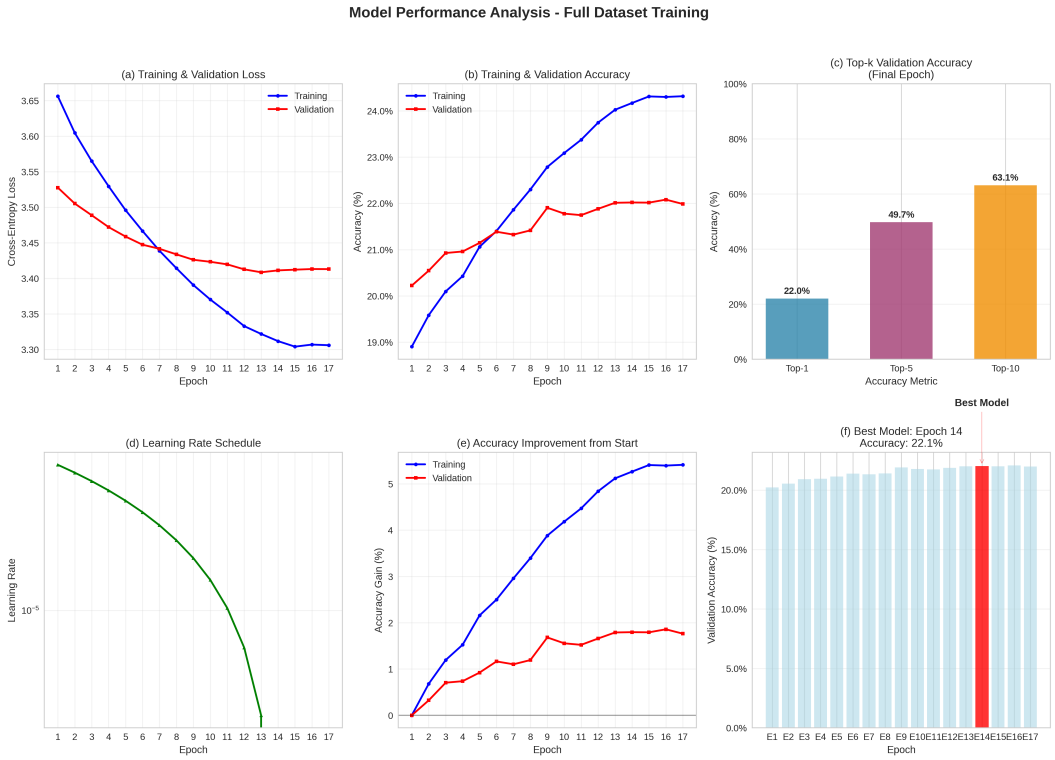


Figure 1: Training performance: (a) Loss curves, (b) Accuracy progression, (c) Top-k accuracy, (d) Learning rate schedule.

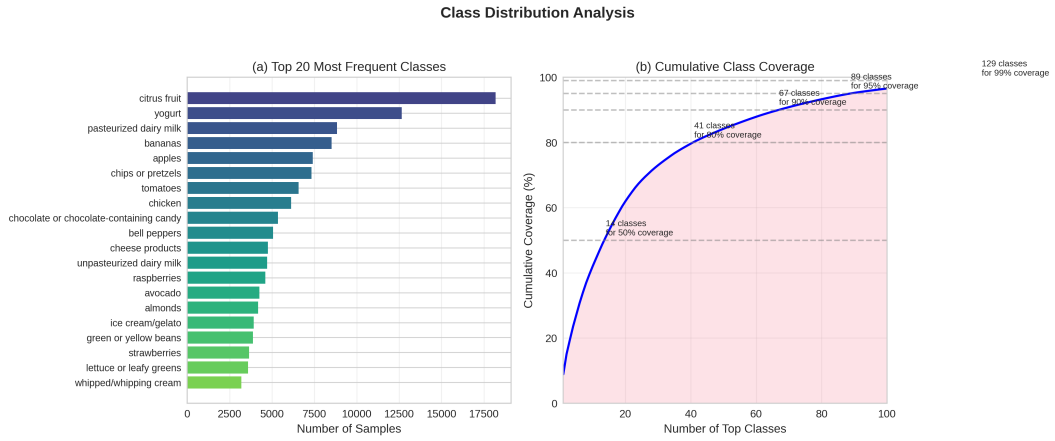


Figure 2: Class distribution: (a) Top 20 frequent classes, (b) Cumulative coverage.

#### 4.2.3 Prediction Analysis Examples

The model demonstrated meaningful prediction distributions. For example, in a sequence dominated by apples and milk, it correctly predicted "pasteurized dairy milk" with 88.8% confidence. For more model outputs, please refer to GitHub.

#### 4.2.4 Top-k Performance

The model's top-5 accuracy (49.70%) more than doubled the top-1 accuracy, and top-10 accuracy reached 63.11%. This indicates strong performance in identifying relevant items, however, exact category prediction is challenging with 182 classes.

## 5 Methodology and Results (Fast Food Category Prediction)

### 5.1 Methodology

#### 5.1.1 Feature Engineering

A separate hybrid approach model will be trained for fast food category prediction using 3,586 survey samples across 12 fast food categories. The binary features (survey questions "Yes": 1, otherwise: 0) are transformed into grocery order-like sequences compatible with our pre-trained BERT model. All demographic data has been excluded, and only the questions the individual answered about food consumed in the past 7 days have been included.

#### 5.1.2 Model Architectures

Six distinct model configurations are evaluated with different feature representations, and architectures are tested:

1. **Logistic Regression (Binary Features)**: Traditional logistic regression model using only the 107 binary food consumption data as input. This will be the baseline model.
2. **Random Forest (Binary Features)**: Traditional random forest decision tree model on binary food consumption features.
3. **XGBoost (Binary Features)**: Standard XGBoost model with binary features.
4. **XGBoost (BERT Embeddings)**: XGBoost trained on only the 256-dimensional embeddings extracted from the fine-tuned BERT-mini model's [CLS] token.
5. **XGBoost (Derived Features)**: Uses just 9 engineered features from BERT outputs, including prediction entropy ($-\sum p_i \log p_i$), top-3/top-5 confidence scores, processed food score heuristics by classifying binary food categories to processed foods, and embedding statistics.
6. **XGBoost Hybrid (Binary + BERT)**: The main model's architecture, which concatenates all three features: binary features (107 dim), BERT embeddings (256 dim), and derived features (9 dim) into a 372-dimensional feature vector for classification. The three feature vectors are concatenated into a single 372-dimensional representation and classified using XGBoost with 100 trees, maximum depth 5, and a learning rate of 0.1.

#### 5.1.3 Training Process

All models were trained with (80%)-(20%) train-test split, with XGBoost parameters consistent across all setups (100 trees, maximum depth 5, learning rate 0.1). The pre-trained BERT-mini model is only used for inference on the new data to extract embeddings.

#### 5.1.4 Evaluation Metrics

Performance was evaluated using correct classification accuracy, weighted F1-score, and a "not-none" classification metric, which only looks at accuracy and weighted F1-score of classes that are not "none" to analyze the model's ability to predict classes other than the most frequent one (70% of the respondents don't have fast food).

## 5.2 Results

### 5.2.1 Overall Performance Comparison

Table 2 shows that XGBoost on just the binary features achieves the highest overall accuracy (50.56%), outperforming the main model (48.19%). However, the main model demonstrated better performance for detecting fast food preference ("not-none") with 57.1% accuracy and 0.389 F1-score, representing the best balance between identifying consumption preferences and maintaining overall classification performance.

Table 2: Model Performance Comparison for Fast Food Category Prediction

| **Model** | **Features** | **Accuracy** | **F1-Score** | **Not-None Acc.** | **Not-None F1** |
|---|---|---|---|---|---|
| XGBoost | Binary | 0.506 | 0.401 | 0.564 | 0.352 |
| Random Forest | Binary | 0.499 | 0.387 | 0.545 | 0.300 |
| **XGBoost Hybrid (Main)** | **Binary + BERT** | 0.482 | 0.381 | **0.571** | **0.389** |
| Logistic Regression | Binary | 0.478 | 0.363 | 0.540 | 0.304 |
| XGBoost | BERT | 0.460 | 0.309 | 0.479 | 0.079 |
| XGBoost | Derived Features | 0.453 | 0.308 | 0.474 | 0.087 |

### 5.2.2 Feature Importance Analysis

The main model's feature importance analysis shows that BERT-derived features contributed 85.8% of the total importance, while binary features contributed just 14.2%. Among the derived features, standard deviation of prediction probabilities (12.64%) and top-5 confidence scores (12.06%) were the most influential. The most important binary features are `chicken` (1.82%), and `eggs` (1.52%).

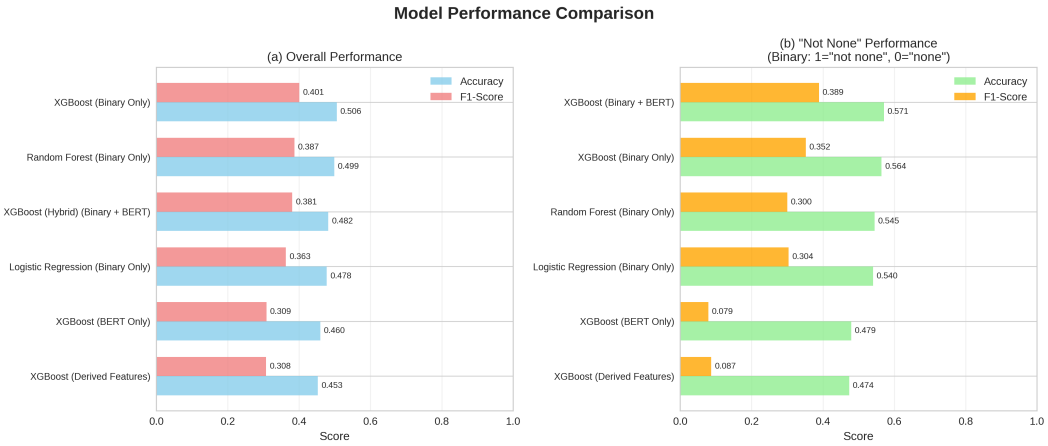


Figure 3: Model Comparison: (a) Standard dataset, (b) "Not-None" dataset.

## 6 Conclusion

The prediction model performs better than random guessing, however, it is still not as great a predictor as income level and other demographics. This project has obvious room for further research. Primarily, not having direct access to fast-food purchase and market data requires the need to use of related datasets from different years and, more importantly, different countries. It is quite likely that the means and variances of these datasets are different, given that they are from separate economic and cultural regions. Further, the microdata only tracks the food consumption of the last 7 days, so any person who did not have fast food in the past week gets labeled as "none" when they might clearly have a preference. Regardless, the model results encourage a deeper look into using grocery data to predict consumer fast food preferences as an alternative to the highly sensitive demographic information of individuals.

## 7 Github Link

https://github.com/ShreyanMal/Food-Services-Market-Analysis

## References


Thomas Burgoine, Chinmoy Sarkar, Crispin J. Webster, and Pablo Monsivais. Examining the interaction of fast-food outlet exposure and income on diet and obesity: evidence from 51,361 UK Biobank participants. *International Journal of Behavioral Nutrition and Physical Activity*, 15(1): 71, 2018. doi: 10.1186/s12966-018-0699-8.

Government of Canada. Foodbook 2.0, public use microdata file 2023. https://open.canada.ca/data/en/dataset/1efcd118-a3df-4cd0-86ae-e4233386b0c6.

Instacart. Instacart Online Grocery Basket Analysis Dataset. https://www.kaggle.com/datasets/yasserh/instacart-online-grocery-basket-analysis-dataset.

Nimit N. Shah and Cheryl D. Fryar and Brian Tsai and Cynthia L. Ogden. Fast Food Intake Among Adults in the United States, August 2021–August 2023. https://www.cdc.gov/nchs/products/databriefs/db533.htm, 2025. NCHS Data Brief No. 533. National Center for Health Statistics.

Emily Seale, Margaret de Groh, and Linda Greene-Finestone. Fast food consumption in adults living in canada: alternative measurement methods, consumption choices, and correlates. *Applied Physiology, Nutrition, and Metabolism*, 48(2):163–171, 2023. doi: 10.1139/apnm-2022-0252. URL https://doi.org/10.1139/apnm-2022-0252. PMID: 36322952.

Statistics Canada. Statistics Canada. Table 11-10-0125-01 Detailed food spending, Canada, regions and provinces. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1110012501, a.

Statistics Canada. Statistics Canada. Table 21-10-0232-01 Food services and drinking places, e-commerce sales. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2110023201&pickMembers%5B0%5D=2.1&cubeTimeFrame.startYear=2016&cubeTimeFrame.endYear=2023&referencePeriods=20160101%2C20230101, b.

Statistics Canada. Statistics Canada. Table 21-10-0171-01 Food services and drinking places, summary statistics. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2110017101, c.