Who Cares? A Predictive Modeling of Child Care in Canada

Mitzie Irene P. Conchada[1]

Karen Ann H. Francisco

[1] Professor, School of Economics, De La Salle University

Table of Contents

1. Executive Summary ..... **Error! Bookmark not defined.**
2. Introduction 5
2.1. Background 5
2.2. Problem Statement 5
2.3. Objectives and Measurement 6
2.4. Assumptions and Limitations 6
3. Data Sources 8
3.1. Data Set Introduction 8
3.2. Exclusions 8
3.3. Initial Data Cleansing or Preparation 9
3.4. Data Dictionary 10
4. Data Exploration 12
4.1. Data Summaries / Descriptive Statistics 12
4.2. Data Visualization 14
4.3. Data Quality 17
5. Data Preparation 19
5.1. Data Preparation needs 19
5.2. Dimension Reduction/ Feature Selection 19
5.3. Correlation Analysis 21
5.4. Dimension Reduction using Classification and Regression Trees 22
6. Model Exploration 24
6.1. Modeling Techniques 24
Classification Trees 24
6.1..1. Full Decision Tree 24
6.1..2. One-Depth Decision Tree 27
6.1..3. Five Depth Decision Tree 28
6.1..4. Full Decision Tree with Cross-Validation 30
6.1..5. GridSearch Cross Validation 30
6.1..6. Random Forest Tree 32
Logistic Regression 33
6.1..7. Full Regression 34
6.1..8. Forward Regression 36

6.1..9. Backward Regression 38
6.1..10. Stepwise Regression 41
6.1..11. Naïve Bayes 43
6.1..12. Neural Networks 44
6.1..13. Support Vector Machine 45
6.2. Sub model for Day Care Type, Child Care Arrangement 46
6.3. Performance Evaluation Metrics 48
7. Model Comparison and Recommendation 50
7.1. Model Selection 50
7.2. Best Model 52
7.3. Model Theory 54
7.4. Model Sensitivity to Key Drivers 55
8. Conclusion and Recommendations 56
8.1. Impacts on Business Problem 56
8.2. Recommended Next Steps 57
References 60
Appendix 62

1. Abstract

The child care sector in Canada has faced significant challenges in recent years, particularly due to increased demand for services amidst limited availability. This study, utilizing data from the *Canadian Survey on Early Learning and Child Care (2023)* conducted by Statistics Canada, investigates the factors influencing parents' or guardians' likelihood of utilizing child care arrangements. The objective is to provide evidence-based insights to inform public policy, improve resource allocation, and enhance service planning.

Predictive analytics and machine learning models, including Classification Tree, Logistic Regression, Naïve Bayes, Neural Networks, and Support Vector Machines, were employed to identify key determinants of child care utilization. The Full Logistic Regression model, with nine variables, achieved the highest performance, yielding an ROC-AUC score of 75%, an F1-score of 72%, and an accuracy rate of 71%.

The findings reveal three primary factors driving child care decisions: dual-income dynamics, specific industry drivers, and the balance between family and work responsibilities. These results underscore the importance of economic and demographic factors in shaping child care needs. The study provides a robust foundation for targeted policy recommendations, advocating for measures that enhance accessibility and affordability in the child care sector.

2. Introduction

2.1. Background

In recent years, the child care sector in Canada has experienced profound changes, driven by evolving societal needs and demographic shifts. Access to affordable and reliable child care has becoming a growing concern for many families, particularly in the wake of the Covid-19 pandemic. The increased demand for child care services, coupled with limited availability, has placed considerable pressure on the system, leading to widespread challenges.

In 2023, a growing number of parents and guardians in Canada reported difficulties in securing child care. Statistics Canada (2024) reported an increase from 36% in 2019 to 49% in 2023. This rise has led to longer waitlists and significant impacts on work-life balance, including changes to work or study schedules, reducing working hours, and postponed returns to the workforce. One of the primary factors contributing to these challenges is heightened demand for child care, with a 34% increase in center-based child care compared to pre-pandemic levels of 31%.

In response to these challenges, the government launched the nation-wide Early Learning and Child Care Program in 2021. The program is designed to enhance quality of child care by upholding and advancing evidence-based quality frameworks and standards. It also aims to make child care more affordable, with the goal of reducing fees to an average of $10 a day, and to increase accessibility by creating 250,000 new child care spaces by 2026 (Statistics Canada, 2024).

2.2. Problem Statement

Building on the issues outlined above, this study seeks to answer the questions: What factors determine the likelihood of parents or guardians utilizing child care arrangements? Moreover, how can the results of this study guide public policy in resource allocation and service planning?

The growing demand for child care services highlights the critical need to understand the underlying factors influencing parent's utilization of these arrangements. Despite various public

and private efforts to provide adequate child care options, disparities persist in accessibility and affordability, affecting different segments of the population in diverse ways. Macdonald (2023) stressed that Canadian parents are increasingly aware of the difficulty in finding child care, a challenge that persists despite rapidly falling fees. Lower fees because of the $10 a day mandate are encouraging more parents to consider child care options. However, he mentioned there are some areas that are considered as 'child care deserts', where there are more children competing for every licensed child care spot (para. 3).

2.3. Objectives and Measurement

Given the research questions posted earlier, the study aims to achieve the following objectives:

a. Identify key determinants of child care utilization through predictive analytics – Investigate the factors that influence the likelihood of parents or guardians using child care arrangements, including economic, demographic and geographic variables.
b. Explore the relationship between the various determinants and child care utilization – Identify which variables are strongly correlated with child care utilization and determine the nature (positive or negative) of these correlations.
c. Develop policy recommendations for expanding child care access – Use the study's findings to propose strategies for increasing the availability of child care spaces and improving accessibility for families in need.

In order to achieve these objectives, the study first conducted exploratory data analysis on the variables through descriptive statistics and data visualizations. Second, various machine learning models were utilized to predict the likelihood of attending child care. Finally, the best model was chosen based on the following metrics: accuracy, Receiver Operating Characteristic - Area under the Curve (ROC-AUC), and F1-Score.

2.4. Assumptions and Limitations

The study relies on several assumptions to ensure the validity of its findings. It assumes that the machine learning models employed are appropriate and results are valid within the study's

context. Moreover, it assumes stability in economic and policy conditions related to child care, as well as the relevance and sufficiency of the selected variables in capturing factors influencing child care utilization.

However, the study considered several limitations. The accuracy of the study's findings depends on the quality of data, which may be complete or inaccurate. Additionally, the results may not generalize to other regions or populations with different socio-economic conditions or child care infrastructures. The study assumes a stable child care landscape, but changes in policies related to child care, economic conditions, or demographic shifts could affect the results and their applicability in the future. Machine learning models, while useful, have limitations such as potential overfitting or underfitting, and the accuracy and interpretability of predictions depend on model selection and tuning. Finally, unobserved or unmeasured variables may also impact the results, leading to an incomplete understanding of child care utilization. These assumptions and limitations should be taken into account when interpreting the study's results and applying them to policy recommendations.

3. Data Sources

3.1. Data Set Introduction

The study used data from Statistics Canada (2024), Canadian Survey on Early Learning and Child Care 2023. The survey, the first after the implementation of the program, was conducted from January to June 2023 across 10 provinces in Canada. A stratified sampling method was employed to select a sample of 29,718 observations among parents and guardians on early learning and child care arrangements for children aged 0 to 5. The survey included information on child care arrangements, including costs, challenges in finding child care, and parents' perception on work and family responsibilities. It also examined the relationship between parents' labor market participation and their use of child care. All variables in the survey were categorical variables and represented by numeric values.

The data set was already cleaned by Statistics Canada (2024) to address missing, invalid or inconsistent data by assigning replacement values to create consistent records. For instance, personal-level and household income data were imputed for missing values, often due to respondent refusal or lack of knowledge. Statistics Canada (2024) employed the nearest neighbor imputation. Errors in questionnaire flow were addressed by marking skipped questions with specific codes. "Valid skip" codes (6, 96, or 996) were used when skips were based on answer questions, while "Not stated" codes (9, 99, or 999) were applied for skips due to non-response. Remaining empty items were filled with numeric codes reserved for indicating "Not stated" for processing purposes (Statistics Canada, 2024).

3.2. Exclusions

Since the study focused only on the demand side of child care, supply side variables such as early learning and child care arrangements (typical day of the week attended, evening/weekend attended, care type choice, difficulty in finding child care, reasons for not using, hours per week, type of arrangements, impact of difficulties and language used in child care) were not included.

Moreover, survey specific variables were dropped such as public use microdata file identifier (ID) and weight used for analysis.

3.3. Initial Data Cleansing or Preparation

Initial data cleansing is a critical step to ensure that data is clean, consistent and ready for analysis. The following steps were implemented:

- ***Data collection and importing***: Data on Canadian Survey on Early Learning and Child Care 2023 the was downloaded from Statistics Canada. The CSV file was imported into the working environment, Python.
- ***Understanding the data structure***: The data was inspected using the various functions in Python to get a sense of the data's structure, types of variables and basic statistics. The dataset was comprised of categorical variables represented by numeric values, with 29,718 rows (observations) and 167 columns (variables).
- ***Handling missing data***: The data set did not have any missing values since this was already addressed by Statistics Canada. To check this, the study employed the command df.isnull().sum()
- ***Dealing with outliers:*** In order to identify outliers, the study employed visualization particularly box plots. The figure below shows an example of a boxplot for respondent's age. There are outliers for the Valid skip and Not stated. The study kept these outliers as their meaning that could potentially affect the results.

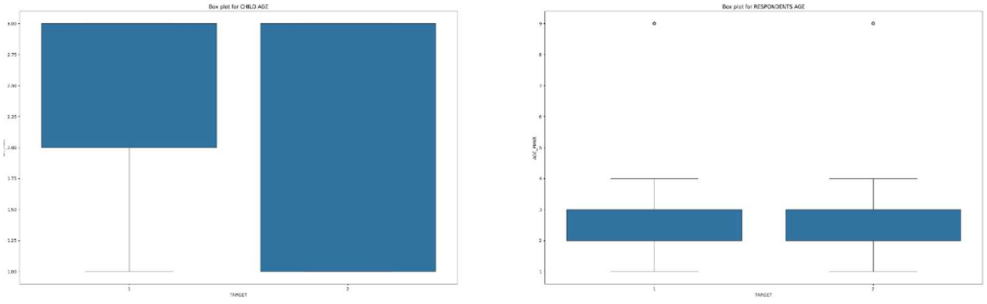

Figure 1. Box Plot for Age

- ***Data transformation:*** Since all categorical variables were numeric, conversion was not necessary except for the target variable ARR_05 (attended child care in past 3 months). The target variable was converted into binary 0 and 1.
- ***Data balancing:*** The original data set was imbalanced with 56% who responded yes to child care and 44%. Though there is a slight difference and this difference reflects the sample at the time of the survey, they study balanced the data through random selection. Downsampling was implemented towards the respondents who chose child care through random selection. Balancing the data set based on the target variable helps in preventing bias towards the majority class. Moreover, it helps improve model performance. Logistic regression models try to maximize the likelihood of the observed data, which can be heavily skewed by imbalanced data.
- ***Initial variable selection:*** The study focused on demand side variables, thus only the following were selected – derived variables (mostly demographic variables of the household), education, citizenship, labor market activities, and perception on balancing work and child care.

3.4. Data Dictionary

The table below shows each of the target and demand side variables (31 variables) along with its description:

Count	Variable	Description
Target variable		
1	TARGET	Enrolled in child care in past 3 months
Demographic variables		
2	AGE_PMKR	Age of respondent (grouped)
3	AGE_SPR	Age of spouse (grouped)
4	CH_AGE	Age of child (grouped)
5	ED_PMKR	Highest educational attainment of respondent
6	ED_SPR	Highest educational attainment of spouse
7	HHINCR	Household income (grouped)
8	IID_HH	Household indigenous group
9	IM_HH	Immigration to Canada, 5 years
10	M_DIFFR	Main difficulty finding child care

11	NCHILD_R	Number of children that live in household (aged 0-17)
12	PBP_HH	Household born in Canada
13	PROV	Province
14	SPFLAGR	Respondent has spouse/partner in household
15	VISMINHH	Household visible minority
16	ALLCOSTG	Annual cost of child care (grouped)
17	NUMCARER	Number of child care arrangements
18	SUBSFL_R	Child receives a subsidy
Employment and labor market activities		
19	IND2_PR	Industry sector of respondent
20	IND2_SR	Industry sector of spouse
21	LMA601PR	Part time or full time work - respondent
22	LMA601SR	Part time or full time work - spouse
23	LM_PMKR	Employed status - respondent (employed or not employed)
24	LM_SPR	Employed status - spouse (employed or not employed)
25	MA_PMKR	Main activity - respondent
26	MA_SPR	Main activity - spouse
27	OCC1_PR	Occupation broad category - respondent
28	OCC1_SR	Occupation broad category - spouse
Balancing work and child care		
29	WLB_01R	Difficulty to fulfill family responsibilities because of work
30	WLB_02R	Difficulty to fulfill work because of family responsibilities
31	WLB_03R	Satisfaction with the balance between job and home life

Table 1 Data Dictionary

4. Data Exploration

Data Exploration is essential for understanding the data before performing detailed analysis and modeling. It is a critical step where we explore and understand the global landscape of the data: structure, patterns and relationships. The purpose is to discover patterns into the data that can help in further analysis, model building, and decision-making. The insights gained from data exploration can significantly improve the quality of the analysis and models, leading to more accurate and interpretable results.

4.1. Data Summaries / Descriptive Statistics

Descriptive statistics provide numerical summaries of the data such as mean, median, mode, minimum, maximum, standard deviation, and quartiles that are helpful in getting familiar on the characteristics of each variable.

Figure below shows the summary of statistics for all the 166 variables excluding the target ARR_05.

Data Summary

dataframe	Values
Number of rows	29718
Number of columns	167

Data Types

Column Type	Count
int64	166
float64	1

number

column_name	NA	NA %	mean	sd	p0	p25	p50	p75	p100	hist
PUMFID	0	0	64380	35780	2438	33660	64050	95320	126800	
WTPP	0	0	75.48	62.5	10	19.02	71.47	109.8	510	
NCHILD_R	0	0	2.161	1.42	1	1	2	3	9	
SPFLAGR	0	0	1.88	0.3245	1	2	2	2	2	
AGE_PMKR	0	0	3.356	2.238	1	2	3	3	9	
AGE_SPR	0	0	4.01	2.363	2	2	3	6	9	
CH_AGE	0	0	1.962	0.825	1	1	2	3	3	
PROV	0	0	33.17	16.24	10	13	35	47	59	
ARR_05	0	0	1.432	0.4953	1	1	1	2	2	
ARR_10AR	0	0	3.388	2.319	1	1	2	6	9	
ARR_10BR	0	0	3.606	2.123	1	2	2	6	9	
ARR_10DR	0	0	3.639	2.09	1	2	2	6	9	
ARR_10FR	0	0	3.697	2.028	1	2	2	6	9	
ARR_10ER	0	0	3.693	2.033	1	2	2	6	9	
NUMCARER	0	0	3.24	2.432	1	1	2	6	9	
ARRT10AR	0	0	5.201	1.259	1	4	6	6	9	
ARRT10BR	0	0	5.568	1.241	1	6	6	6	9	
ARRT10DR	0	0	5.775	0.7573	1	6	6	6	9	
ARRT10ER	0	0	5.854	0.7553	1	6	6	6	9	
ARRT10FR	0	0	5.897	0.6592	1	6	6	6	9	

Figure 2. Sample Descriptive Statistics

In this specific study, all predictors are categorical in nature; therefore, frequency counts, unique values, and the top values are more valuable to help understand the characteristics of each variable. A modified summary of these values are seen below (30 variables):

No.	Variable	Description	Unique	Top	Freq
Demographic variables					
1	AGE_PMKR	Age of respondent (grouped)	5	2	11176
2	AGE_SPR	Age of spouse (grouped)	5	3	11190
3	ALLCOSTG	Annual cost of child care (grouped)	6	6	12827
4	CH_AGE	Age of child (grouped)	3	1	9788
5	ED_PMKR	Highest educational attainment of respondent	4	3	12350
6	ED_SPR	Highest educational attainment of spouse	5	3	8409
7	HHINCR	Household income (grouped)	8	7	7159
8	IID_HH	Household indigenous group	3	2	23813
9	IM_HH	Immigration to Canada, 5 years	3	2	21924
10	NCHILD_R	Number of children that live in household (aged 0-17)	4	2	11374
11	NUMCARER	Number of child care arrangements	4	6	12827
12	M_DIFFR	Main difficulty finding child care	12	96	15447
13	MA_PMKR	Main activity - respondent	5	1	12520
14	MA_SPR	Main activity - spouse	6	1	19325
15	PBP_HH	Household born in Canada	3	2	16208
16	PROV	Province	10	35	4473
17	SPFLAGR	Respondent has spouse/partner in household	2	2	22614
18	SUBSFL_R	Child receives a subsidy	5	6	12654
19	VISMINHH	Household visible minority	3	2	16838
Employment and labor market activities					
20	LM_PMKR	Employed status - respondent (employed or not employed)	3	1	18467
21	IND2_PR	Industry sector of respondent	21	96	7082
22	IND2_SR	Industry sector of spouse	21	99	6153
23	LM_SPR	Employed status - spouse (employed or not employed)	4	1	20543
24	LMA601PR	Part time or full time work - respondent	4	2	14912
25	LMA601SR	Part time or full time work - spouse	4	2	19397
26	OCC1_PR	Occupation broad category - respondent	11	96	7082
27	OCC1_SR	Occupation broad category - spouse	11	8	5792
Balancing work and child care					
28	WLB_01R	Difficulty to fulfill family responsibilities because of work	7	6	7082
29	WLB_02R	Difficulty to fulfill work because of family responsibilities	7	6	7082
30	WLB_03R	Satisfaction with the balance between job and home life	7	6	7082

Table 2. Descriptive Statistics (30 variables)

4.2. Data Visualization

Visualization techniques help in understanding data easily using images, graphs, charts and helps in identifying trends, relationships, and outliers. It is a powerful tool to explore the multidimensional nature of the dataset and an effective way to present the results graphically

without the need to delve into raw numbers. The following basic charts and distribution plot are used to help in getting familiar with the data structure.

- Histogram: shows the distribution of a single variable and how data points are spread across different values. Looking at the distribution plot, particularly for household income, it can be seen that it does not have a non-linear relationship.

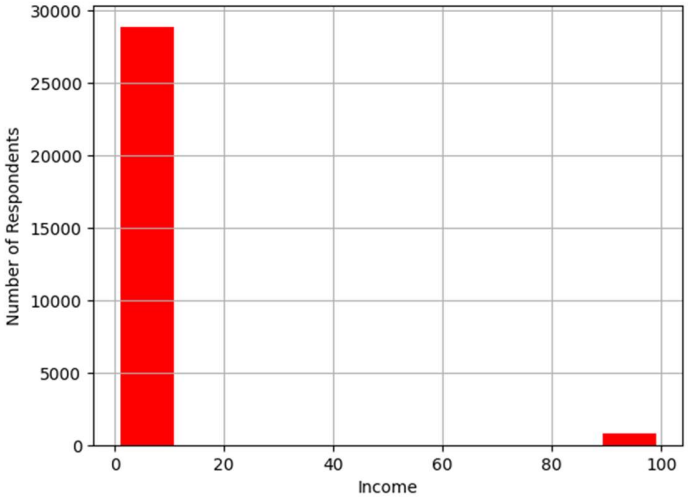

Figure 3. Household Income Histogram

- Boxplot: reveals the entire distribution of a numerical variable and identifies outliers in the data. Although boxplots are typically used for numerical data, for this study, boxplot with category counts have been created to check for outliers. Any category count that lies outside the whiskers can be considered an outlier.

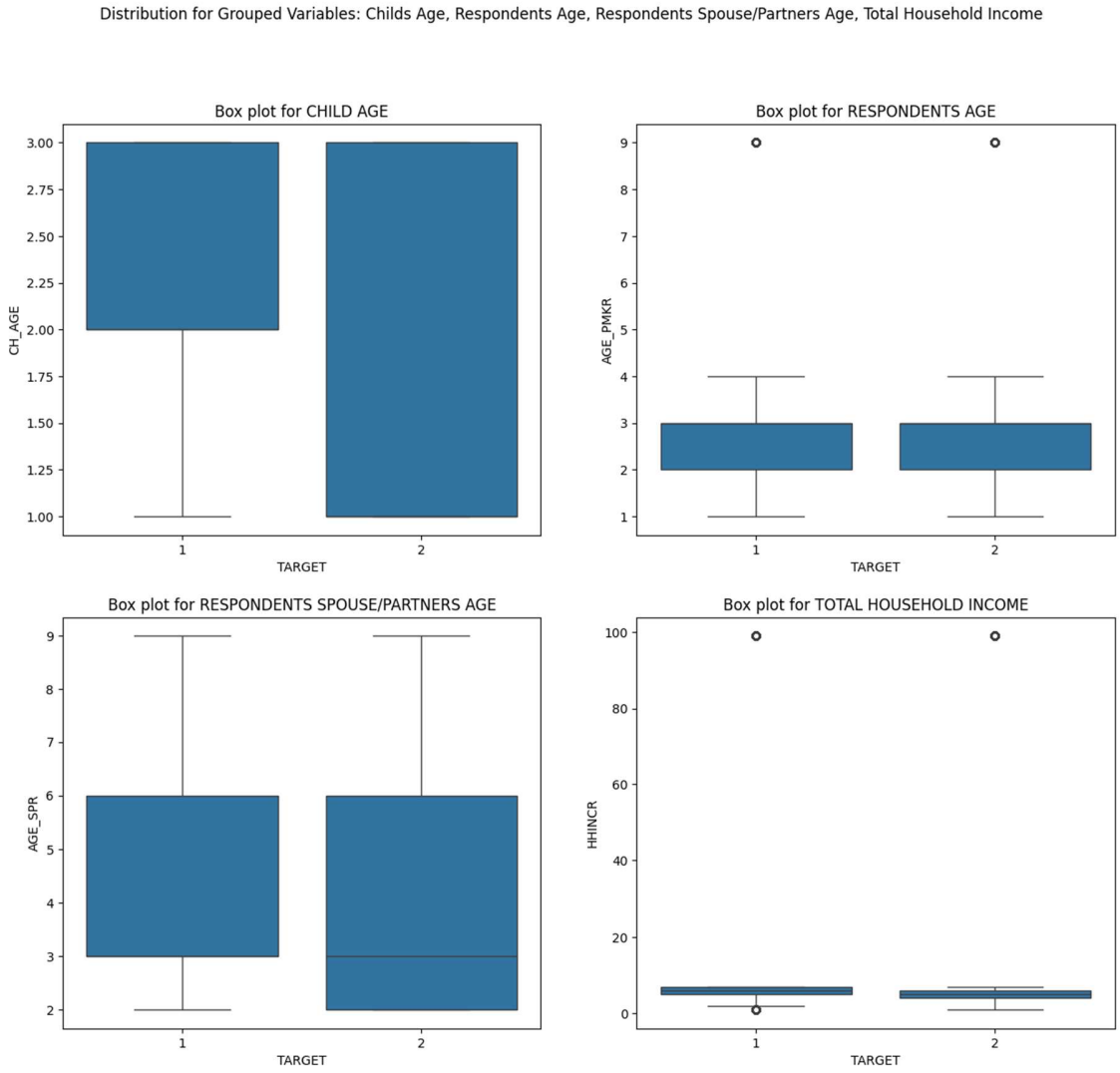

Figure 4 Boxplot

- Scatterplot: illustrates relationships between two continuous variables. Because this study uses only categorical variables and a basic scatter plot must be numerical variables, displaying one is not necessary.

- Heatmap: helps in understanding relationship between multiple variables. Since the dataset deals with categorical variables, the heatmap visualizes the frequency of occurrences in a contingency table (cross-tabulation of two categorical variables).

Figure below reveals the variables with strong positive or negative correlation based on the correlation analysis used for dimension reduction.

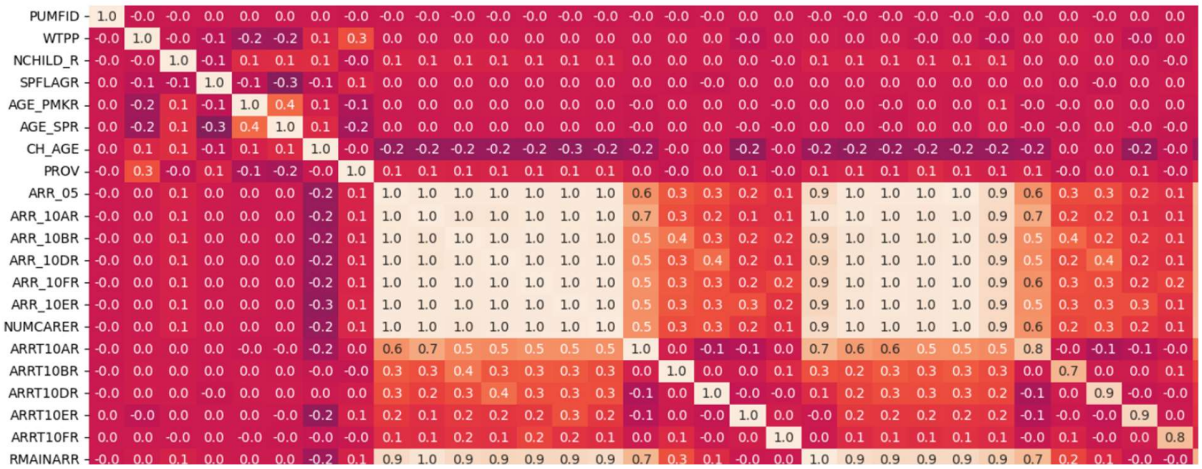

Figure 5. Sample Heatmap for all 166 variables

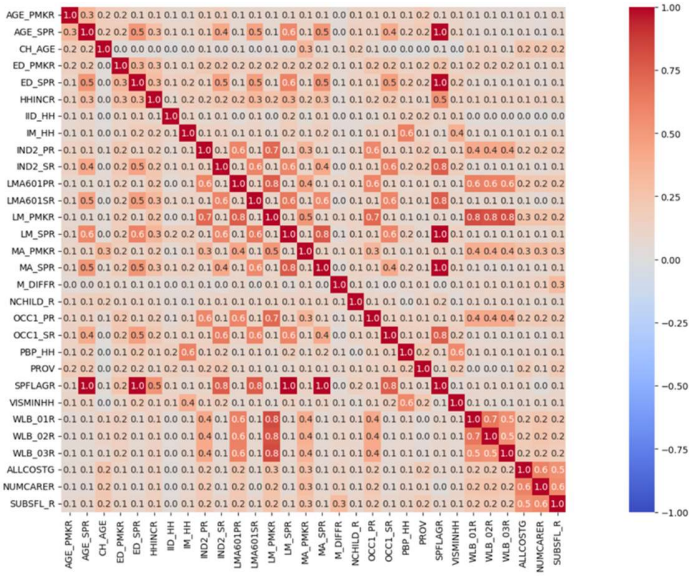

Figure 6. Heatmap for 30 variables (valid columns)

4.3. Data Quality

Data quality involves examining the data for its data type, structure, and relationships. It is important to recognize that Statistics Canada's data sets have undergone rigorous data cleaning

processes to ensure high quality and reliability, making them suitable for robust analysis. These processes include correcting inconsistencies, handling missing data, and standardizing formats, which are critical steps in preparing the data for accurate and meaningful statistical analysis. Details on this are also discussed in the governance paper.

Nevertheless, the study still checked for the following:

- Missing Values: No missing values were identified in the data set
- Duplicates: No duplicate records were found in the data set
- Data Types: All variables are coded as integer types in raw data including the target variables
- Target Class Distribution: Imbalanced target class distribution having 57% use child care and 43% did not use child care

	Count	Percentage
TARGET		
1	16891	56.83761
2	12827	43.16239

Table 3. Frequency Table for Target Variable (attended child care or not)

5. Data Preparation

5.1. Data Preparation needs

- Imputation: No imputation was needed as there were no missing values.
- Transformation: No need for transformation as data was non-skewed or no outliers were detected.
- Data Type Conversion: Data type converted for predictors and target variable.
- Predictors: converted from integer (int) type to object type as it is categorical variable.
- Target: converted into binary format (1: attended child care in past 3 months, 0: did not attend child care in past 3 months)
- Random Sampling: Based on the target distribution, there is an imbalanced dataset that could lead to biased performances, inaccurate results and overfitting; thus, random sampling is employed before modeling. Balanced dataset contains 25,654 observations
- Data Partition: A 60-40 data partition split is used, 60% for training the model, and the remaining 40% is for testing or validation. This is a balanced approach given the size of the dataset and enough for the model to learn the data while providing a validation set for evaluating performance and detecting overfitting.

5.2. Dimension Reduction/ Feature Selection

Dimension reduction is a technique used in eliminating number of input variables in a dataset that are not needed. In this study, we are working with large number of variables in the dataset and this technique is crucial to simplify the model, improve the performance and reduces the computation cost. Several dimension reduction approaches were used to avoid overfitting and produce more robust results before deploying a model.

- **Valid Columns based on domain knowledge and business problem**

  Out of 166 variables in the raw data set, only 30 variables have been identified as relevant to the outcome of interest based on the logic and domain knowledge. The table below shows the variables that are deemed necessary for the focus of this study. All

selected variables pertain to the demand side of child care. This includes Household characteristics, Labor market activities and Perception in balancing work and child care.

Count	Variable	Description
Target variable		
1	TARGET	Enrolled in child care in past 3 months
Demographic variables		
2	AGE_PMKR	Age of respondent (grouped)
3	AGE_SPR	Age of spouse (grouped)
4	CH_AGE	Age of child (grouped)
5	ED_PMKR	Highest educational attainment of respondent
6	ED_SPR	Highest educational attainment of spouse
7	HHINCR	Household income (grouped)
8	IID_HH	Household indigenous group
9	IM_HH	Immigration to Canada, 5 years
10	M_DIFFR	Main difficulty finding child care
11	NCHILD_R	Number of children that live in household (aged 0-17)
12	PBP_HH	Household born in Canada
13	PROV	Province
14	SPFLAGR	Respondent has spouse/partner in household
15	VISMINHH	Household visible minority
16	ALLCOSTG	Annual cost of child care (grouped)
17	NUMCARER	Number of child care arrangements
18	SUBSFL_R	Child receives a subsidy
Employment and labor market activities		
19	IND2_PR	Industry sector of respondent
20	IND2_SR	Industry sector of spouse
21	LMA601PR	Part time or full time work - respondent
22	LMA601SR	Part time or full time work - spouse
23	LM_PMKR	Employed status - respondent (employed or not employed)
24	LM_SPR	Employed status - spouse (employed or not employed)
25	MA_PMKR	Main activity - respondent
26	MA_SPR	Main activity - spouse
27	OCC1_PR	Occupation broad category - respondent
28	OCC1_SR	Occupation broad category - spouse
Balancing work and child care		
29	WLB_01R	Difficulty to fulfill family responsibilities because of work
30	WLB_02R	Difficulty to fulfill work because of family responsibilities
31	WLB_03R	Satisfaction with the balance between job and home life

Table 4. Valid Columns based on Domain Knowledge

- **Irrelevant Columns**

  There are 136 variables that are considered to be irrelevant given the nature of the research problem and the goal of the analysis. All variables related to supply side in child care arrangements have been removed.

5.3. Correlation Analysis

In a large dataset, it is likely that there will be overlapping information from all the predictors. To avoid redundancies and duplicate variables in this study, a correlation matrix was used.

From the initial dataset, total of 135 are highly correlated when correlation analysis was performed thus causing overfitting in the model. By incorporating domain knowledge as mentioned earlier, the variables were reduced to 30. To check on the overlap of information in the dataset, statistical tests such as Chi-Square Test and Cramer's V were done to find redundancies on these 30 variables and assess the association and strength of the relationship, if any. These tests are used for categorical variables.

- Chi-Square Test: used to check the association between categorical variables and assess if there is a significant relationship between the variables.
- Cramer's V: a post-test used to measure how strong the association between two categorical variables after performing Chi-Square test.

Based on the correlation matrix, 15 significant variables were identified as highly correlated; these are the variables having value greater than 0.7 strength of association from Cramer's rule. To avoid multicollinearity problems in the model, eliminating variables that are highly correlated to others is important.

```
Variables: IND2_PR and LM_PMKR have a correlation of 0.71
Variables: LMA601SR and SPFLAGR have a correlation of 0.77
Variables: LMA601PR and LM_PMKR have a correlation of 0.76
Variables: AGE_SPR and SPFLAGR have a correlation of 1.00
Variables: LM_PMKR and WLB_01R have a correlation of 0.83
Variables: OCC1_SR and SPFLAGR have a correlation of 0.77
Variables: ED_SPR and SPFLAGR have a correlation of 1.00
Variables: IND2_SR and SPFLAGR have a correlation of 0.77
Variables: MA_SPR and SPFLAGR have a correlation of 1.00
Variables: LM_SPR and MA_SPR have a correlation of 0.77
Variables: LM_PMKR and WLB_03R have a correlation of 0.82
Variables: LM_PMKR and OCC1_PR have a correlation of 0.72
Variables: LM_SPR and SPFLAGR have a correlation of 1.00
Variables: LM_PMKR and WLB_02R have a correlation of 0.80
```

Selected features using Cramer's rule:

No.	Variable	Description
1	AGE_SPR	Age of spouse (grouped)
2	ED_SPR	Highest educational attainment of spouse
3	IND2_PR	Industry sector of respondent
4	IND2_SR	Industry sector of spouse
5	LM_PMKR	Employed status - respondent (employed or not employed)
6	LM_SPR	Employed status - spouse (employed or not employed)
7	LMA601PR	Part time or full time work - respondent
8	LMA601SR	Part time or full time work - spouse
9	MA_SPR	Main activity - spouse
10	OCC1_PR	Occupation broad category - respondent
11	OCC1_SR	Occupation broad category - spouse
12	SPFLAGR	Respondent has spouse/partner in household
13	WLB_01R	Difficulty to fulfill family responsibilities because of work
14	WLB_02R	Difficulty to fulfill work because of family responsibilities
15	WLB_03R	Satisfaction with the balance between job and home life
16	TARGET	Enrolled in child care in past 3 months

Table 5. Selected Features from Cramer's V

5.4. Dimension Reduction using Classification and Regression Trees

One way to remove columns is using an intrinsic method such as tree based algorithms that support feature selection functions. In this case, Random Forest is used to select a subset of most relevant features from the original dataset based on their predictive

power. Total of 9 important features are selected from Random Forest Tree. Threshold is defined as 0.05 for the number of top features to be retained.

feature	importance	std
AGE_SPR	0.057347	0.010327
ED_SPR	0.060671	0.009900
OCC1_PR	0.083953	0.026661
OCC1_SR	0.087349	0.010476
IND2_PR	0.105067	0.016491
WLB_03R	0.109346	0.054180
IND2_SR	0.114373	0.008611
WLB_01R	0.124625	0.068670
WLB_02R	0.130017	0.066729

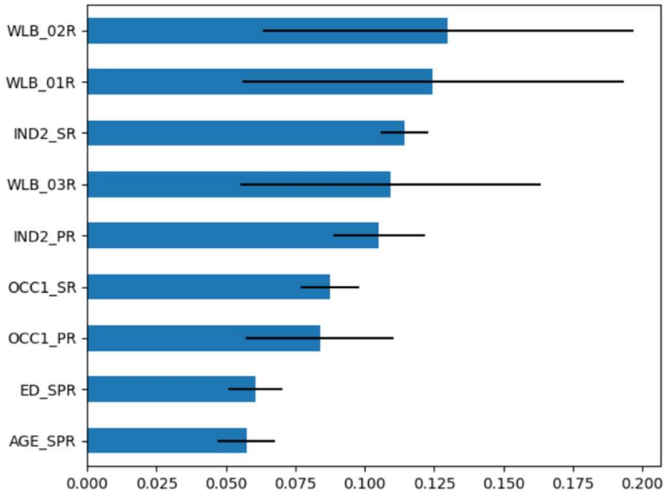

Figure 7. Feature Selection from Random Forest

Summary table for 9 variables is found below.

No.	Variable	Description
1	AGE_SPR	Age of spouse
2	ED_SPR	Highest educational attainment of spouse
3	OCC1_PR	Industry sector of respondent
4	OCC1_SR	Industry sector of spouse
5	IND2_PR	Occupation broad category - respondent
6	IND2_SR	Occupation broad category - spouse
7	WLB_01R	Difficulty to fulfill family responsibilities because of work
8	WLB_02R	Difficulty to fulfill work because of family responsibilities
9	WLB_03R	Satisfaction with the balance between job and home life
10	TARGET	Enrolled in child care in past 3 months

Table 6. Summary of 9 Variables from Random Forest Feature Selection

6. Model Exploration

Model exploration is a pivotal phase in analytics project that involves evaluating various models and this process is essential for aligning the analytical strategy with the research goals.

In this case, the goal is both predictive and explanatory modeling. By systematically testing different models, we can identify the approach that best balances predictive accuracy with interpretability, ensuring that the chosen model not only performs well but also provides valuable insights.

Different models are built and tested in this study to choose the optimal model that drives both understanding and decision-making.

6.1. Modeling Techniques

Classification Trees

Trees are the most widely used model as it easy to interpret among the data-driven methods. It is a tree-like structure that contains two types of node: decision nodes and terminal nodes. Decision nodes represent the decision rule whereas terminal nodes denote the outcome of the class level. Aside from the interpretability advantage of trees, it also captures complex, non-linear relationships between predictors and target variable. However, trees can cause overfitting leading to poor generalization on new data. There is a need to fine-tune parameters to stop the tree growth. We have employed different trees to address the overfitting of data in the modeling.

6.1..1. Full Decision Tree

Full decision tree is used as starting point in the modeling or the baseline. This kind of decision tree has been grown to its full potential without any pruning or controlling the parameters. Initially, the full decision tree is ran using all the 166 predictors in the dataset which yielded 100% accuracy both in training and validation set, so a need for dimension reduction is needed.

In the previous section of this paper, it was mentioned that 30 valid columns were selected based on the knowledge of each variable. We have run another full decision tree using these 30 valid columns and the result is still overfitting. For the discussion of models below, two different number of predictors tested:

First Model: 15 significant variables selected from correlation analysis

No.	Variable	Description
1	AGE_SPR	Age of spouse (grouped)
2	ED_SPR	Highest educational attainment of spouse
3	IND2_PR	Industry sector of respondent
4	IND2_SR	Industry sector of spouse
5	LM_PMKR	Employed status - respondent (employed or not employed)
6	LM_SPR	Employed status - spouse (employed or not employed)
7	LMA601PR	Part time or full time work - respondent
8	LMA601SR	Part time or full time work - spouse
9	MA_SPR	Main activity - spouse
10	OCC1_PR	Occupation broad category - respondent
11	OCC1_SR	Occupation broad category - spouse
12	SPFLAGR	Respondent has spouse/partner in household
13	WLB_01R	Difficulty to fulfill family responsibilities because of work
14	WLB_02R	Difficulty to fulfill work because of family responsibilities
15	WLB_03R	Satisfaction with the balance between job and home life

Table 7. Model 1 with 15 significant variables

Second Model: 9 variable importance selected from random forest

No.	Variable	Description
1	AGE_SPR	Age of spouse
2	ED_SPR	Highest educational attainment of spouse
3	OCC1_PR	Industry sector of respondent
4	OCC1_SR	Industry sector of spouse
5	IND2_PR	Occupation broad category - respondent
6	IND2_SR	Occupation broad category – spouse
7	WLB_01R	Difficulty to fulfill family responsibilities because of work
8	WLB_02R	Difficulty to fulfill work because of family responsibilities
9	WLB_03R	Satisfaction with the balance between job and home life

Table 8. Model 2 with 9 significant variables

Model Result

	Number of splits	Number of terminal nodes	Most important feature
Model 1 (15)	9660	4731	WLB_02R
Model 2 (9)	9696	4849	WLB_02R

Table 9 Full Decision Tree Result Summary

- **Model 1(15 variables)**

Figure 8 Full Decision Tree for Model 1 (15 variables)

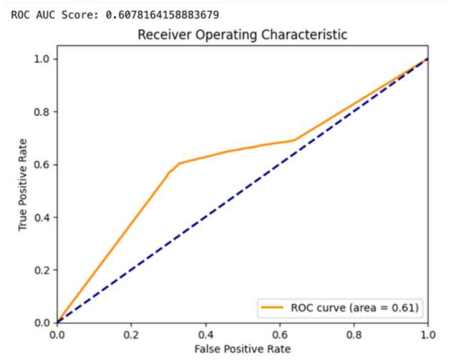

Figure 9 ROC-AUC Full Decision Tree Model 1 (15 variables)

- **Model 2(9 variables)**

Figure 10 Full Decision Tree for Model 2 (9 variables)

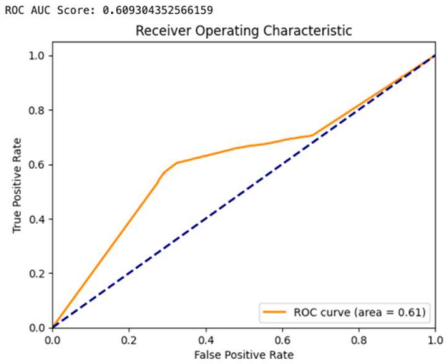

Figure 11 ROC-AUC Full Decision Tree for Model 2 (9 variables)

Interpretation

Since we have large dataset, the resulting tree is too big to visualize and analyze. Both models' important or first split is WLB_02R (Difficulty to fulfill work because of family responsibilities) .

6.1..2. One-Depth Decision Tree

One-depth tree is the simplest form of a decision tree and provides a baseline model for comparison. It gives us a sense of how well a basic model can perform on the given data.

Model Result

	Number of splits	Number of terminal nodes	Most important feature
Model 1 (15) & Model 2 (9)	2	2	WLB_02R

Table 10 One-Depth Tree Result Summary

- **Model 1 (15) and Model 2 (9) show same result**

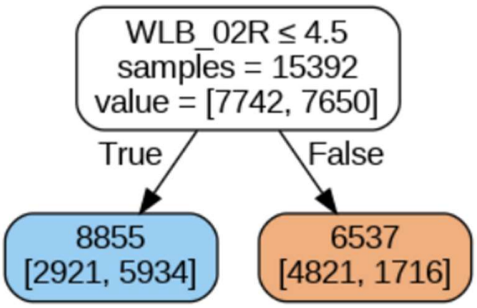

Figure 13 One-Depth Tree for Model 1(15) and Model 2(9)

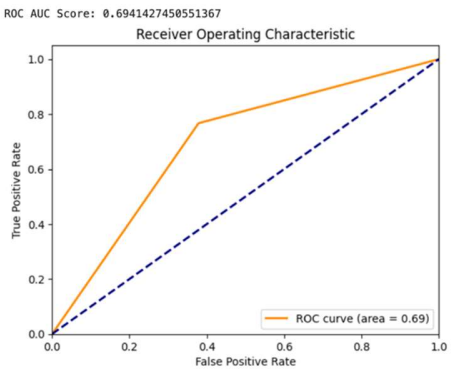

Figure 12 ROC-AUC One-Depth Tree for Model 1(15) and Model 2(9)

Interpretation

A one-depth tree revealed that WLB_02R (Difficulty to fulfill work because of family responsibilities) is the most critical factor in making decision. This information is valuable as it serves as a guide in the development of more sophisticated models.

6.1..3. Five Depth Decision Tree

A five-depth tree is a decision tree with a maximum depth of five levels. This captures a more complex relationships within the data compared to a one-depth tree.

Model Result

	Number of splits	Number of terminal nodes	Most important feature
Model 1 (15) and Model 2 (9)	62	32	WLB_02R

Table 11 Five-Depth Tree Result Summary

- **Model 1(15)**

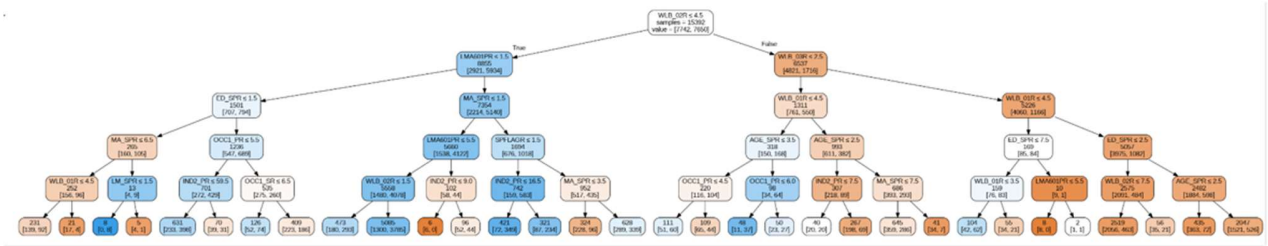

Figure 14 Five-Depth Tree for Model 1 (15)

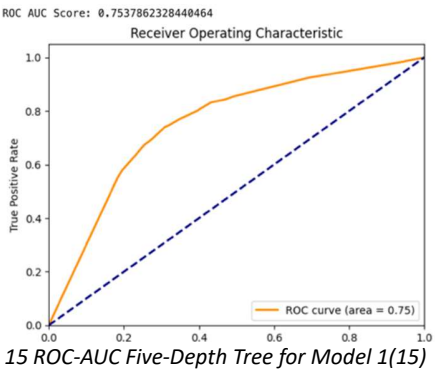

Figure 15 ROC-AUC Five-Depth Tree for Model 1(15)

- **Model 2(9)**

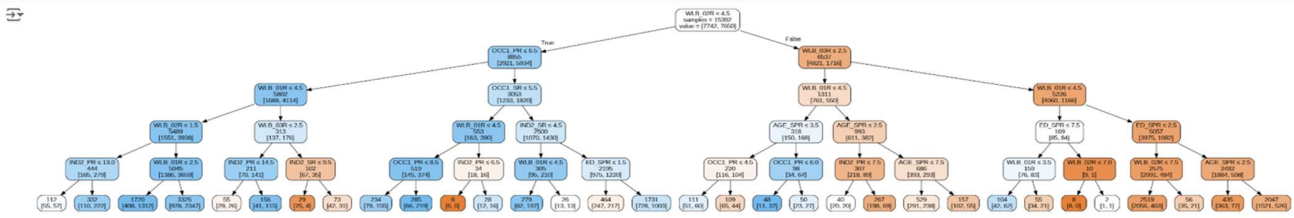

Figure 16 Five-Depth Tree for Model 2(9)

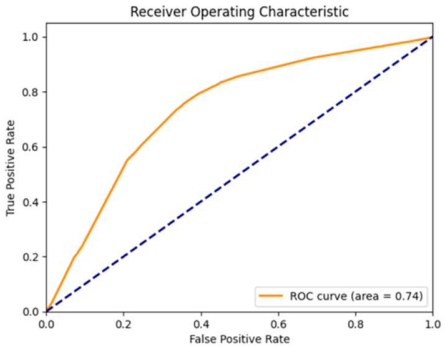

Figure 17 ROC-AUC Five-Depth Tree for Model 2(9)

Interpretation

- ***Model 1 (15)***

All those experiencing difficulty to fulfill work responsibilities because of family wherein the primary respondent works part time, and the spouse/partner highest educational attainment is high school diploma are less likely to use child care arrangements.

- ***Model 2 (9)***

A five-depth tree showed 5 important interactions between features. All those experiencing difficulty to fulfil work responsibilities because of family, and where the primary respondents work at senior management positions. Moreover, it included industry specifics like Mining, Utilities, Construction, Manufacturing, Wholesale trade, Finance, Information, Real Estate and Professional Services are more likely to use child care arrangements.

6.1..4. Full Decision Tree with Cross-Validation

This tree is an implementation of a full decision tree with 5-fold cross-validation to evaluate how well this highly complex model generalizes to unseen data

Model Result

The following are the scores of each fold, which indicate the performance of the model on the validation set.

- ***Model 1(15)***

1st Fold	62.93%
2nd Fold	63.91%
3rd Fold	64.38%
4th Fold	62.96%
5th Fold	61.70%

Table 12 Score of Each Fold (Model 1(15))

- ***Model 2(15)***

1st Fold	60.98%
2nd Fold	62.88%
3rd Fold	61.21%
4th Fold	62.43%
5th Fold	59.02%

Table 13 Score of Each Fold (Model 1(15))

Interpretation

Based on the results, accuracy varied across the different folds for both models.

6.1..5. GridSearch Cross Validation

GridSearch Cross Validation is one way to optimize the tree parameters. As mentioned earlier, trees can cause overfitting and we need to stop the tree growth. We have performed this by controlling the maximum depth of the tree. The GridSearch Cross Validation is a powerful method to find the optimal hyperparameters for a model without manually trying each

combination. In this model, we set cross-validation to five (5). The hyperparameters is selected as optimal set if it gives the best average performance across the 5 folds.

Model Result

- ***Model 1(15)***

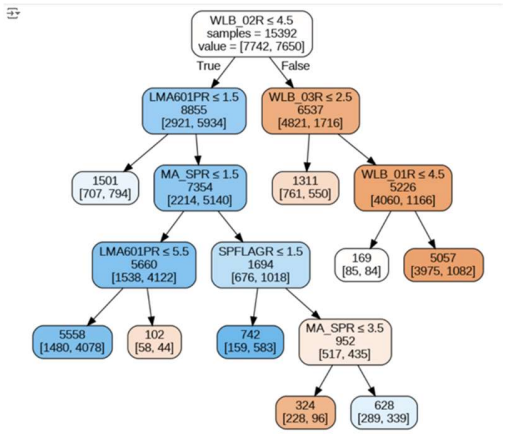

Figure 18 GridSearch CV for Model 1(15)

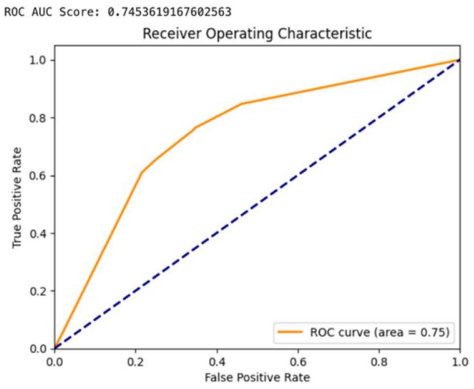

Figure 19 ROC-AUC GridSearch CV for Model 1(15)

- ***Model 2 (9)***

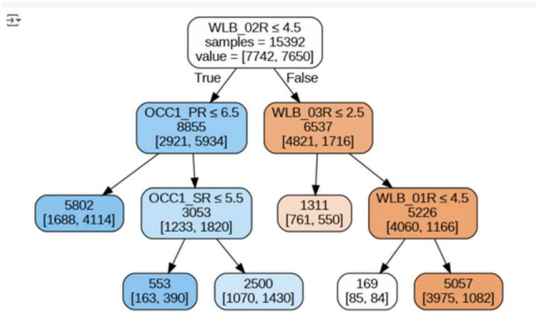

Figure 20 GridSearch CB for Model 2(9)

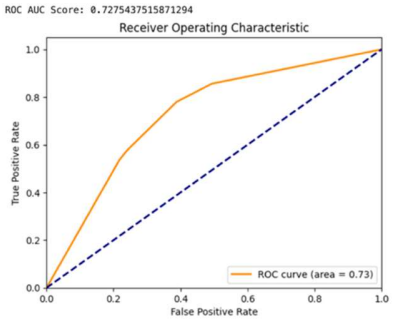

Figure 21 ROC-AUC GridSearch CV for Model 2(9)

Interpretation

- ***Model 1 (15)***

All those experiencing Difficulty to fulfill work responsibilities because of family or fulfill family responsibilities because of work are more likely to use child care arrangements.

- ***Model 2(9)***

All those experiencing Difficulty to fulfill work responsibilities because of family and primary respondent is working are more likely to use child care arrangements.

6.1..6. Random Forest Tree

Random forest tree is an ensemble method that consists of multiple decision trees and combined to improve predictive performance. In the early section, this method was used as one of the techniques for dimension reduction. Given that our goal is both predictive and explanatory, random forest is also utilized as the one of the predictive models. Moreover, this model is known for its high predictive power and ability handle complex relationship in the data.

Model Result

- ***Model 1 (15)***

	feature	importance	std
3	SPFLAGR	0.003503	0.003470
4	LM_SPR	0.007826	0.003399
5	LMA601SR	0.017991	0.003287
2	LM_PMKR	0.021512	0.048508
0	MA_SPR	0.030971	0.004292
8	LMA601PR	0.051423	0.047692
12	AGE_SPR	0.057204	0.010264
14	ED_SPR	0.060702	0.010122
11	OCC1_PR	0.083847	0.024042
7	OCC1_SR	0.087717	0.009752
1	IND2_PR	0.105568	0.018650
9	WLB_03R	0.106593	0.050595
6	IND2_SR	0.113439	0.009266
13	WLB_01R	0.116986	0.062396
10	WLB_02R	0.134719	0.069099

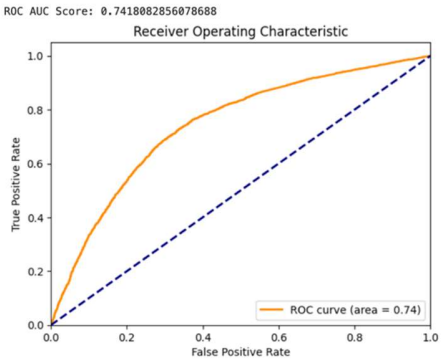

Figure 22 ROC-AUC Random Forest for Model 1(15)

- ***Model 2(9)***

	feature	importance	std
0	AGE_SPR	0.063531	0.011938
1	ED_SPR	0.068079	0.010609
2	OCC1_PR	0.093059	0.031667
3	OCC1_SR	0.101181	0.011923
4	IND2_PR	0.118483	0.020242
5	WLB_03R	0.125013	0.059376
6	IND2_SR	0.131679	0.010135
8	WLB_02R	0.149195	0.079448
7	WLB_01R	0.149781	0.081628

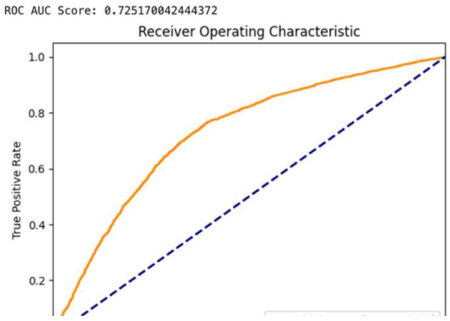

Figure 23 ROC-AUC Random Forest for Model 2(9)

Interpretation

Both models identified 9 variable importance such as:

AGE_SPR	Age of spouse (grouped)
ED_SPR	Highest educational attainment of spouse
IND2_PR	Industry sector of respondent
IND2_SR	Industry sector of spouse
OCC1_PR	Occupation broad category - respondent
OCC1_SR	Occupation broad category - spouse
WLB_01R	Difficulty to fulfill family responsibilities because of work
WLB_02R	Difficulty to fulfill work because of family responsibilities
WLB_03R	Satisfaction with the balance between job and home life

Logistic Regression

Logistic Regression is used for binary classification tasks, where the outcome variable is categorical. is best in predicting and explaining predictors with the outcome. Moreover, the result is easy to understand and interpret. The coefficients of the model indicate the relationship between features and odds of the outcome and used for interpretation in each of the feature importance. The model takes two steps: First is to use the logistic function to map the predicted values to probabilities belonging to each class, then using a cutoff value to classify in the classes. The probability cutoff is set to 0.5. If the value is higher than 0.5, it is classified as 1 (use of child care), otherwise 0 (did not use child care).

Different types of logistic regression such as Full, Forward, Backward and Stepwise logistic regression have been built to see the performance and check if any can be used for feature selection.

6.1..7. Full Regression

Model Result

- **Model 1(15)**

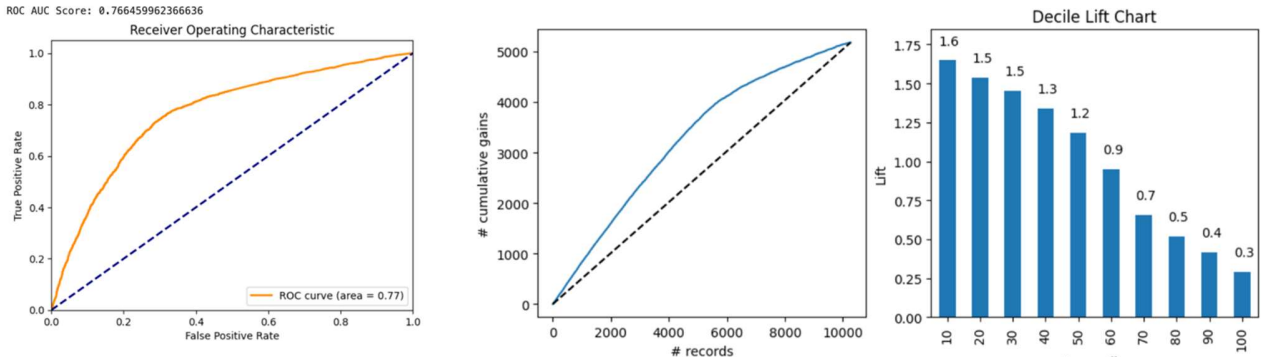

Figure 24 ROC-AUC, Cumulative Gains, Decile Lift Chart for Model 1(15)

- **Model 2(9)**

No.	Variable	Description	coef	odds	p-value
1	WLB_02R_3	Difficulty to fulfil work responsibilities because of family (sometimes)	0.607488	1.835815	0.00E+00
2	WLB_01R_5	Difficulty to fulfill family responsibilities bec of work (never)	-0.91976	0.398616	8.32E-130
3	IND2_SR_12	Industry sector - spouse (professional, scientific & technical)	0.569745	1.767817	5.00E-44
4	IND2_SR_4	Industry sector - spouse (construction)	0.656657	1.928335	1.15E-43
5	WLB_02R_4	Difficulty to fulfil work responsibilities because of family (rarely)	0.548413	1.730505	4.13E-34
6	WLB_02R_2	Difficulty to fulfil work responsibilities because of family (often)	0.593174	1.809724	1.31E-28
7	IND2_SR_20	Industry sector - spouse (public admin)	0.491124	1.634153	8.41E-17
8	IND2_SR_5	Industry sector - spouse (manufacturing)	0.523616	1.68812	9.96E-15
9	IND2_SR_7	Industry sector - spouse (retail trade)	0.561966	1.754117	1.60E-13
10	IND2_SR_6	Industry sector - spouse (wholesale)	0.720791	2.056058	3.53E-12
11	WLB_01R_4	Difficulty to fulfill family responsibilities bec of work (rarely)	-0.30144	0.739753	9.26E-12
12	IND2_SR_3	Industry sector - spouse (utilities)	1.046958	2.84897	2.77E-09
13	IND2_SR_15	Industry sector - spouse (educ services)	0.560067	1.75079	3.92E-09
14	WLB_03R_3	Satisfaction with the balance between job and home life (sometimes)	-0.33855	0.7128	1.89E-07
15	IND2_SR_8	Industry sector - spouse (transportation)	0.335496	1.398634	3.45E-07

16	IND2_SR_19	Industry sector - spouse (other services)	0.506031	1.658695	3.71E-07
17	IND2_SR_11	Industry sector - spouse (real estate)	0.762111	2.142795	2.89E-06
18	IND2_SR_14	Industry sector - spouse (administrative & support, waste mngmt)	0.566497	1.762084	6.98E-06
19	WLB_03R_9	Satisfaction with the balance between job and home life (not stated)	-0.78464	0.456285	1.03E-04
20	IND2_SR_9	Industry sector - spouse (information & cultural industries)	0.549356	1.732137	1.61E-04
21	IND2_SR_16	Industry sector - spouse (health care)	0.322178	1.380131	7.91E-04
22	IND2_SR_2	Industry sector - spouse (mining, quarrying)	0.365194	1.440794	9.24E-04
23	WLB_01R_9	Difficulty to fulfill family responsibilities bec of work (not stated)	-0.70502	0.4941	3.36E-03
24	WLB_01R_2	Difficulty to fulfill family responsibilities bec of work (often)	0.107461	1.113448	1.71E-02
25	IND2_SR_17	Industry sector - spouse (arts, entertainment)	0.451508	1.57068	2.08E-02
26	WLB_02R_5	Difficulty to fulfil work responsibilities because of family (never)	-0.14365	0.866187	2.40E-02
27	IND2_SR_18	Industry sector - spouse (accom & food services)	0.279334	1.322249	4.31E-02

Table 14 Significant Variables with Odds Ratio Based on P-value

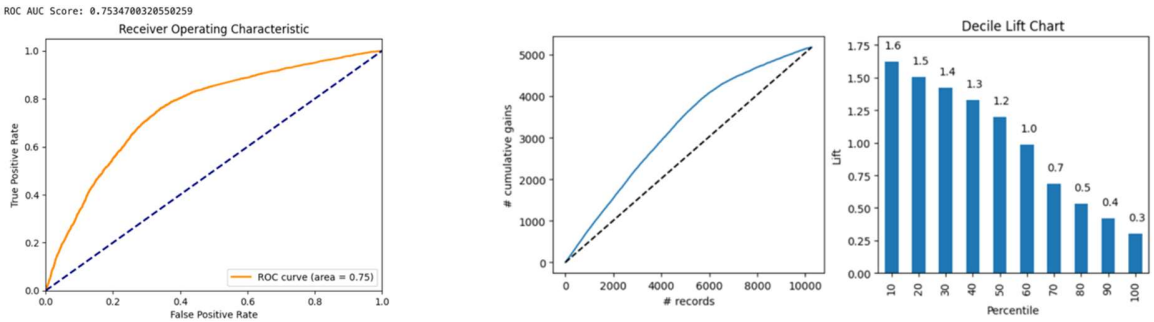

Figure 25 ROC-AUC, Cumulative Gains, Decile Lift Chart for Model 2(9)

6.1..8. Forward Regression

Forward Selection is a method where the model starts with no predictors and progressively adds them to the model. The process continues until adding more predictors does not significantly improve the model.

Model Result

- ***Model 1(15)***

Selected features:

AGE_SPR_4	Age of spouse (grouped)
AGE_SPR_9	Age of spouse (grouped)
IND2_PR_16	Industry sector of respondent
IND2_PR_18	Industry sector of respondent
IND2_PR_8	Industry sector of respondent
IND2_SR_11	Industry sector of spouse
IND2_SR_3	Industry sector of spouse
LM_PMKR_9	Industry sector of spouse
LM_SPR_9	Industry sector of spouse
LMA601PR_2	Part time or full time work - respondent
MA_SPR_2	Main difficulty finding child care
MA_SPR_3	Main difficulty finding child care
MA_SPR_9	Main difficulty finding child care
WLB_01R_3	Difficulty to fulfill family responsibilities because of work
WLB_01R_4	Difficulty to fulfill family responsibilities because of work
WLB_01R_5	Difficulty to fulfill family responsibilities because of work
WLB_02R_2	Difficulty to fulfill work because of family responsibilities
WLB_02R_4	Enrolled in child care in past 3 months
WLB_03R_9	Satisfaction with the balance between job and home life

Table 15. Selected Features form Forward Regression

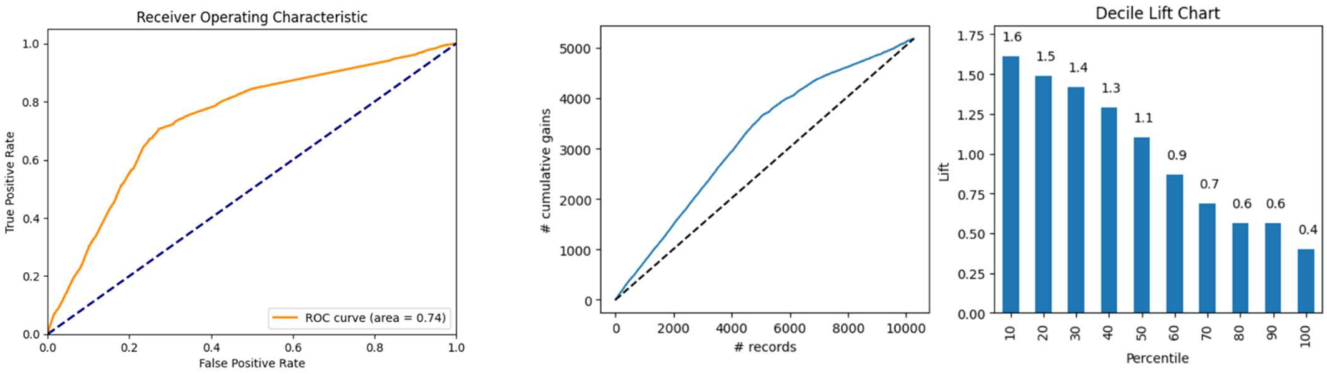

Figure 26 ROC-AUC, Cumulative Gains, Decile Lift Chart Model 1(5)

- ***Model 2(9)***

  Selected features:

AGE_SPR_6	Age of spouse (grouped)
IND2_PR_17	Industry sector of respondent
IND2_PR_18	Industry sector of respondent
IND2_PR_99	Industry sector of respondent
IND2_SR_10	Industry sector of spouse
IND2_SR_18	Industry sector of spouse
OCC1_PR_3	Occupation broad category - respondent
OCC1_PR_7	Occupation broad category - respondent
OCC1_PR_96	Occupation broad category - respondent
OCC1_PR_99	Occupation broad category - respondent
OCC1_SR_96	Occupation broad category - spouse
OCC1_SR_99	Occupation broad category - spouse
WLB_01R_5	Difficulty to fulfill family responsiblities because of work
WLB_01R_9	Difficulty to fulfill family responsiblities because of work
WLB_02R_5	Difficulty to fulfill work because of family responsiblities

Table 16. Selected Features form Forward Regression

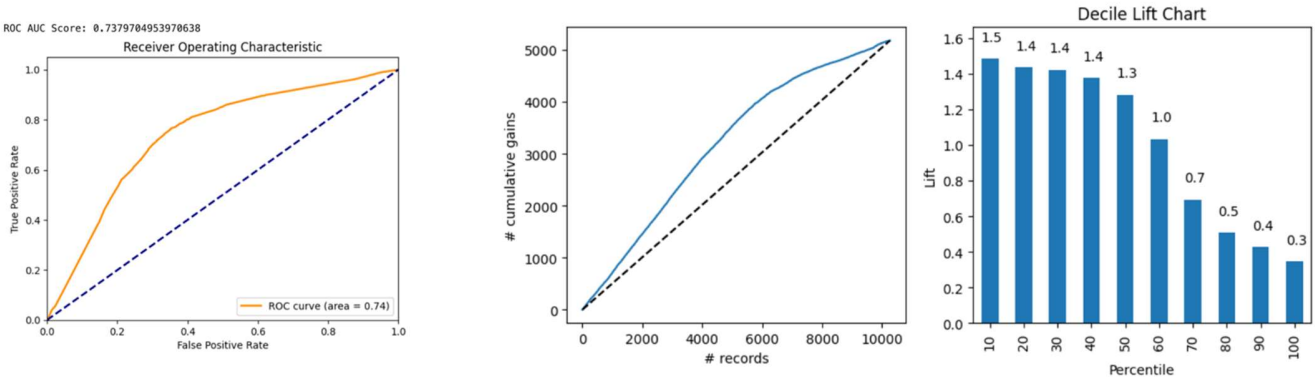

Figure 27 ROC-AUC, Cumulative Gains, Decile Lift Chart for Model 2(9)

6.1..9. Backward Regression

Backward Elimination is a method where you start with all potential predictors and remove them one at a time. The process continues until further removal of predictors does not significantly degrade the model's performance.

Model Result

- ***Model 1(15)***

  Selected features:

AGE_SPR_3	Age of spouse (grouped)
AGE_SPR_4	Age of spouse (grouped)
AGE_SPR_9	Age of spouse (grouped)
ED_SPR_2	Highest educational attainment of spouse
ED_SPR_3	Highest educational attainment of spouse
IND2_PR_11	Industry sector of respondent
IND2_PR_12	Industry sector of respondent
IND2_PR_14	Industry sector of respondent
IND2_PR_16	Industry sector of respondent
IND2_PR_18	Industry sector of respondent
IND2_PR_3	Industry sector of respondent
IND2_PR_4	Industry sector of respondent
IND2_PR_7	Industry sector of respondent
IND2_PR_8	Industry sector of respondent

IND2_PR_99	Industry sector of respondent
IND2_SR_3	Industry sector of spouse
LM_SPR_2	Industry sector of spouse
LM_SPR_9	Industry sector of spouse
LMA601PR_9	Part time or full time work - respondent
LMA601SR_9	Employed status - spouse (employed or not employed)
MA_SPR_2	Main difficulty finding child care
MA_SPR_3	Main difficulty finding child care
MA_SPR_4	Main difficulty finding child care
MA_SPR_9	Main difficulty finding child care
OCC1_PR_10	Occupation broad category - respondent
OCC1_PR_7	Occupation broad category - respondent
OCC1_PR_96	Occupation broad category - respondent
OCC1_SR_10	Occupation broad category - spouse
OCC1_SR_5	Occupation broad category - spouse
OCC1_SR_96	Occupation broad category - spouse
WLB_01R_2	Enrolled in child care in past 3 months
WLB_01R_3	Difficulty to fulfill family responsibilities because of work
WLB_01R_4	Difficulty to fulfill family responsibilities because of work
WLB_02R_2	Difficulty to fulfill work because of family responsibilities
WLB_02R_3	Difficulty to fulfill work because of family responsibilities
WLB_02R_4	Difficulty to fulfill work because of family responsibilities
WLB_02R_5	Difficulty to fulfill work because of family responsibilities
WLB_03R_3	Satisfaction with the balance between job and home life
WLB_03R_9	Satisfaction with the balance between job and home life

Table 17. Selected Features from Backward Regression

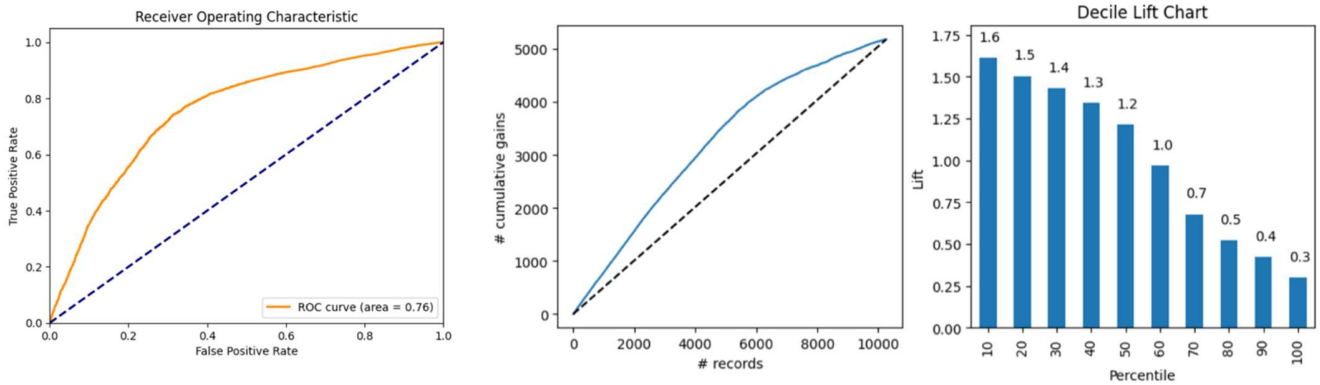

Figure 28 ROC-AUC, Cumulative Gains, Decile Lift Chart for Model 1(15)

- ***Model 2(9)***

Selected features:

AGE_SPR_3	Age of spouse (grouped)
AGE_SPR_4	Age of spouse (grouped)
AGE_SPR_6	Age of spouse (grouped)
AGE_SPR_9	Age of spouse (grouped)
ED_SPR_2	Highest educational attainment of spouse
ED_SPR_3	Highest educational attainment of spouse
OCC1_PR_7	Occupation broad category - respondent
OCC1_PR_10	Occupation broad category - respondent
OCC1_PR_96	Occupation broad category - respondent
OCC1_PR_99	Occupation broad category - respondent
OCC1_SR_4	Occupation broad category - spouse
OCC1_SR_5	Occupation broad category - spouse
OCC1_SR_9	Occupation broad category - spouse
OCC1_SR_10	Occupation broad category - spouse
IND2_PR_3	Industry sector of respondent
IND2_PR_4	Industry sector of respondent
IND2_PR_7	Industry sector of respondent
IND2_PR_8	Industry sector of respondent
IND2_PR_11	Industry sector of respondent
IND2_PR_12	Industry sector of respondent
IND2_PR_14	Industry sector of respondent
IND2_PR_16	Industry sector of respondent
IND2_PR_18	Industry sector of respondent
IND2_PR_99	Industry sector of respondent
WLB_03R_3	Enrolled in child care in past 3 months
WLB_03R_9	Enrolled in child care in past 3 months
IND2_SR_2	Industry sector of spouse
IND2_SR_3	Industry sector of spouse
IND2_SR_4	Industry sector of spouse
IND2_SR_5	Industry sector of spouse
IND2_SR_6	Industry sector of spouse
IND2_SR_7	Industry sector of spouse
IND2_SR_8	Industry sector of spouse
IND2_SR_9	Industry sector of spouse
IND2_SR_10	Industry sector of spouse
IND2_SR_11	Industry sector of spouse
IND2_SR_12	Industry sector of spouse
IND2_SR_14	Industry sector of spouse
IND2_SR_15	Industry sector of spouse
IND2_SR_19	Industry sector of spouse

IND2_SR_20	Industry sector of spouse
IND2_SR_99	Industry sector of spouse
WLB_01R_3	Difficulty to fulfill family responsibilities because of work
WLB_01R_4	Difficulty to fulfill family responsibilities because of work
WLB_01R_5	Difficulty to fulfill family responsibilities because of work
WLB_02R_2	Difficulty to fulfill work because of family responsibilities
WLB_02R_3	Difficulty to fulfill work because of family responsibilities
WLB_02R_4	Difficulty to fulfill work because of family responsibilities

Table 18. Selected Features from Backward Regression

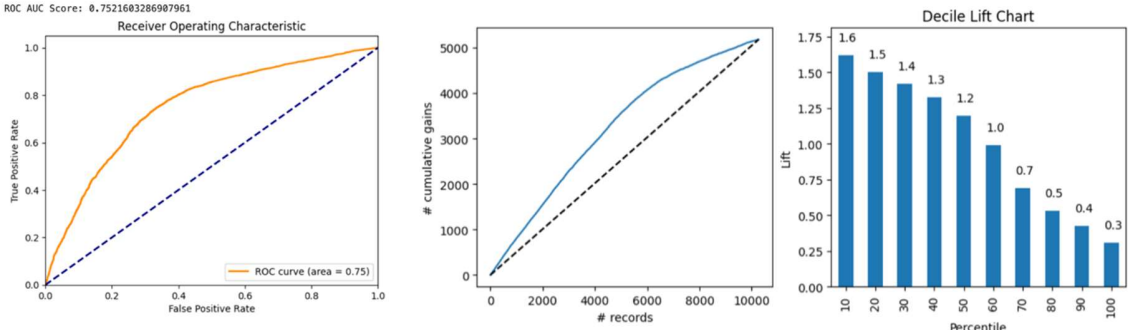

Figure 29 ROC-AUC, Cumulative Gains, Decile Lift Chart for Model 2(9)

6.1..10. Stepwise Regression

Stepwise Selection combines both forward and backward approaches. It involves adding and removing variables iteratively.

Model Result

- ***Model 1(15)***

  Selected features:

IND2_PR_18	Industry sector of respondent
LM_SPR_2	Industry sector of spouse
MA_SPR_2	Main difficulty finding child care
MA_SPR_3	Main difficulty finding child care
MA_SPR_9	Main difficulty finding child care
OCC1_PR_7	Occupation broad category - respondent
OCC1_PR_96	Occupation broad category - respondent
WLB_02R_5	Difficulty to fulfill work because of family responsibilities
WLB_02R_9	Difficulty to fulfill work because of family responsibilities

Table 19. Selected Features from Stepwise Regression

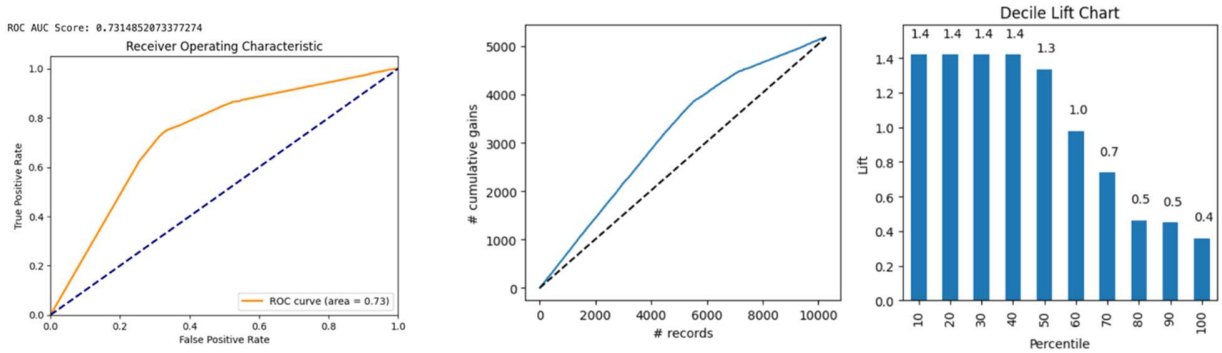

Figure 30 ROC-AUC, Cumulative Gains, Decile Lift Chart for Model 1(15)

- ***Model 2(9)***

  Selected features:

AGE_SPR_6	Age of spouse (grouped)
IND2_PR_18	Industry sector of respondent
OCC1_PR_7	Occupation broad category - respondent
OCC1_PR_96	Occupation broad category - respondent
OCC1_PR_99	Occupation broad category - respondent
WLB_01R_5	Difficulty to fulfill family responsibilities because of work
WLB_03R_9	Satisfaction with the balance between job and home life

Table 20. Selected Features from Stepwise Regression

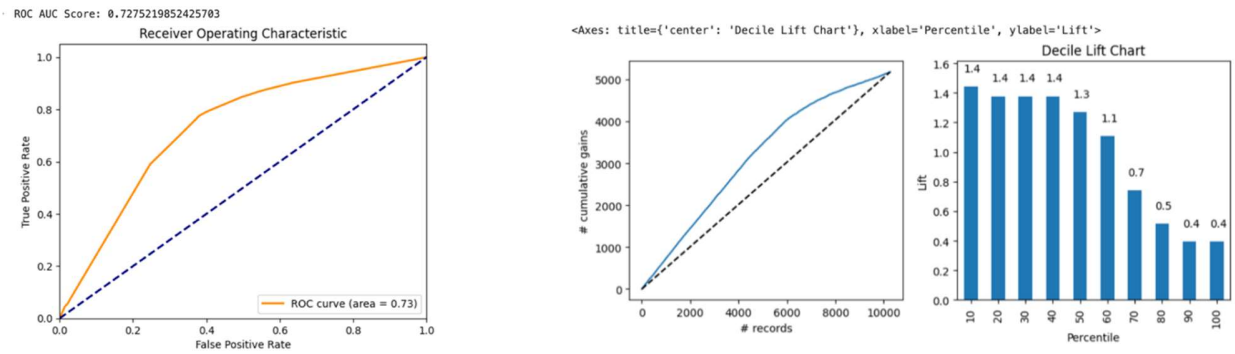

Figure 31 ROC-AUC, Cumulative Gains, Decile Lift Chart for Model 2(9)

6.1..11. Naïve Bayes

Naïve Bayes Classifier is a model based on Bayes' Theorem with a simplifying assumption of conditional independence among features. It has the ability to handle categorical variables directly. A Multinomial Naïve Bayes has been employed as it is deemed appropriate for the features as it represent the counts or frequencies.

Model Result

- **Model 1(15)**

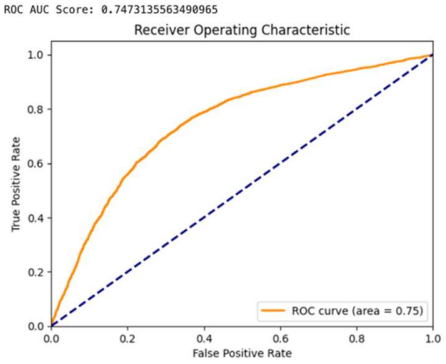

Figure 32 ROC-AUC for Model 1(15)

- **Model 2(9)**

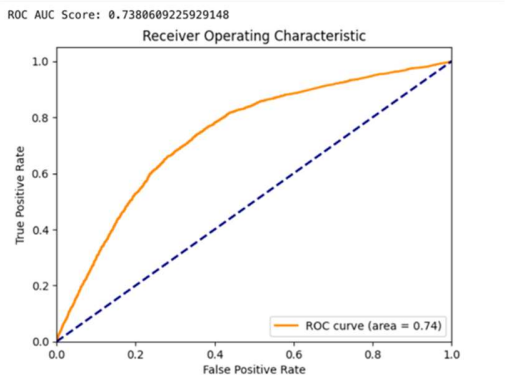

Figure 33 ROC-AUC for Model 2(9)

Interpretation

Model 1(15) has a slightly accuracy score of 74.73% compared to Model 2(9) with 73.86%. This implies that Model 1(15) has a higher accuracy of classifying the results correctly.

6.1..12. Neural Networks

Neural networks are a robust tool for binary classification tasks like predicting child care usage. They excel in modeling complex and non-linear relationships, However, main drawback is the interpretability; it is it considered “black boxes” as it faces interpretability challenges compared to other models. Multi-layer Perceptron (MLP) for classification tasks is used in this model as it good in capturing complex, non-linear relationship in the data.

Model Result

- ***Model 1(15)***

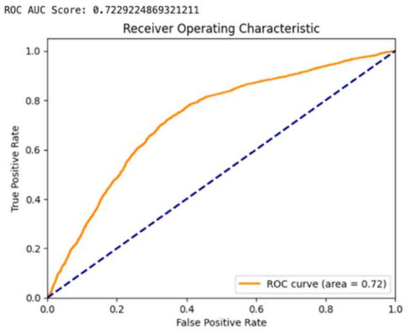

Figure 34 ROC-AUC for Model 1(15)

- ***Model 2(9)***

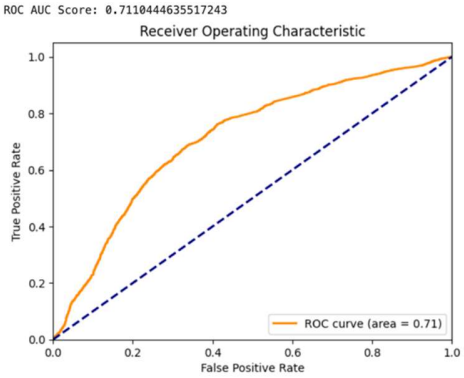

Figure 35 ROC-AUC for Model 2(9)

Interpretation

Model 1(15) has a slightly accuracy score of 72.29% compared to Model 2(9) with 71.10%. This implies that Model 1(15) has a higher accuracy of classifying the results correctly.

6.1..13. Support Vector Machine

Support Vector Machine (SVM) is versatile model used for classification task that can handle both linear and non-linear problems through different kernels. It is robust to overfitting, but less interpretable compared to simpler models like logistic regression or decision trees.

Model Result

- **Model 1(15)**

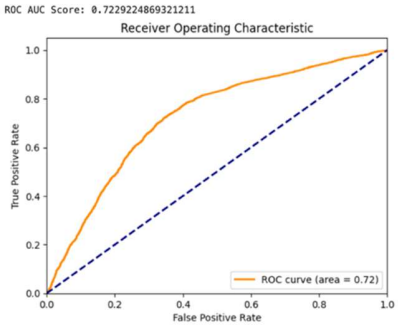

Figure 36 ROC-AUC for Model 1(15)

- ***Model 2(9)***

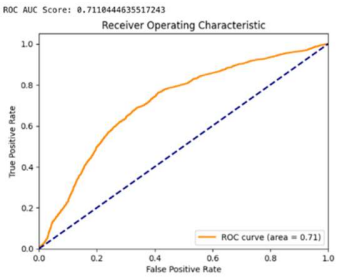

Figure 37 ROC-AUC for Model 2(9)

Interpretation

Model 1(15) has a slightly accuracy score of 72.29% compared to Model 2(9) with 71.10%. This implies that Model 1(15) has a higher accuracy of classifying the results correctly.

6.2. Sub model for Day Care Type, Child Care Arrangement

Sub modeling child care arrangements by focusing specifically on day care provides a nuanced approach to understanding the factors influencing parents' decisions. By examining day care as a distinct category rather than grouping it with general child care use, we aimed to uncover whether specific characteristics or factors are unique to this arrangement. This targeted analysis is crucial for identifying whether different predictors or influences drive the choice of day care over other child care options.

Target is set to ARR_10AR: attended Day Care in the past 3 months

This sub model exploration was tested in all models built for the main child care usage discussed in modeling techniques.

The results show that the same factors are consistently identified across all models, including both general child care use and specific day care arrangements. It suggests that the determinants of child care use are uniformly relevant regardless of the type of arrangement. This finding underscores the robustness of the identified factors and implies that no significant divergence exists between the predictors for general child care use and those for day care specifically.

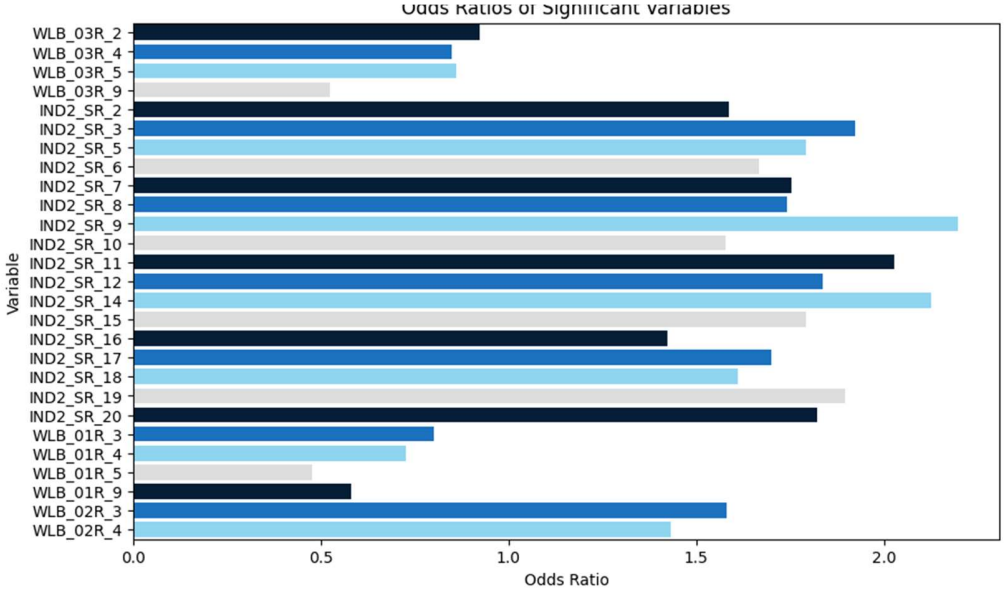

Figure 38 Odds Ratio of Full Logistic Regression RF

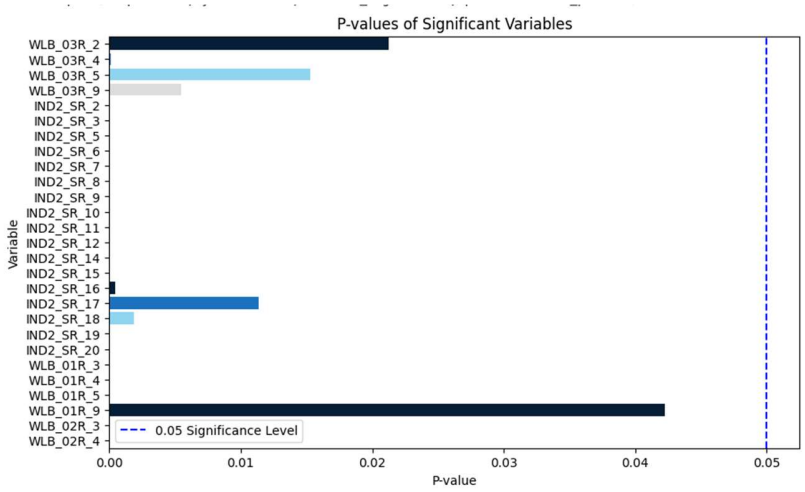

Figure 39 P-value of Full Logistic Regression RF

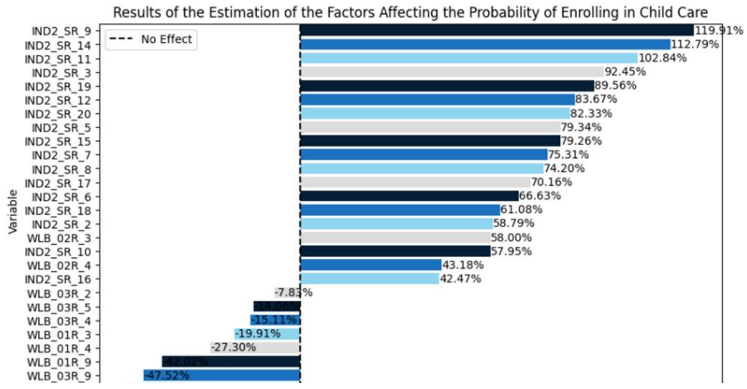

Figure 40 Result of Estimation of Full Logistic Regression RF

Model Name	ROC AUC	Accuracy	Recall	Precision	F1 Score
Backward Logistic Regression	0.69	0.65	0.71	0.62	0.66
SVM	0.68	0.65	0.73	0.62	0.67
Forward Logistic Regression	0.67	0.65	0.70	0.63	0.66
Stepwise Logistic Regression	0.66	0.65	0.70	0.63	0.66
Full Logistic Regression	0.69	0.64	0.72	0.62	0.67
Five-Depth Decision Tree	0.67	0.64	0.72	0.61	0.66
Forward Logistic Regression RF	0.66	0.64	0.77	0.60	0.68
GridSearch Cross Validation	0.66	0.64	0.69	0.62	0.65
Full Logistic Regression RF	0.67	0.63	0.72	0.60	0.66
Backward Logistic Regression RF	0.67	0.63	0.72	0.60	0.66
Five-Depth Decision Tree RF	0.66	0.63	0.75	0.60	0.67
SVM RF	0.66	0.63	0.76	0.60	0.67
Naive Bayes RF	0.66	0.63	0.71	0.60	0.65
GridSearch Cross Validation RF	0.65	0.63	0.78	0.59	0.67
Stepwise Logistic Regression RF	0.64	0.63	0.77	0.59	0.67
One-Depth Decision Tree RF	0.63	0.63	0.77	0.59	0.67
One-Depth Decision Tree	0.63	0.63	0.77	0.59	0.67
Naive Bayes	0.65	0.61	0.67	0.59	0.63
Random Forest	0.65	0.61	0.65	0.60	0.62
Neural Network	0.63	0.61	0.63	0.60	0.61
Random Forest RF	0.63	0.60	0.63	0.58	0.61
Neural Network RF	0.63	0.59	0.60	0.58	0.59
Full Decision Tree RF	0.56	0.57	0.52	0.57	0.54
Full Decision Tree	0.55	0.57	0.53	0.57	0.55

Figure 41 Model Comparison for Day Care

Based on the results in the figure above, the best model still turned out to be the Full Logistic Regression mode with variables from Random Forest. The significant variables that affected child care are very similar to the overall model on child care: occupation type, industry type, and perception on work and family responsibilities.

6.3. Performance Evaluation Metrics

Performance evaluation metrics are used to evaluate how effective each type of models are based on the accuracy, specificity and sensitivity, and reliability.

The following metrics are used to evaluate the models:

- F1-Score: balance between precision and recall in a single metric that balances trade-off between them. In this study, correctly identifying positives while minimizing false positives and false negatives are of great importance. This metric also works well with imbalanced datasets.

- ROC-AUC: area under the ROC curve that shows the graph between False Positive Rate and True Positive Rate. The metric works well with binary classification and less affected by the imbalanced distribution.
- Accuracy: ratio of number of correct predictions to the total number of input instances. We want our model to provide accurate results all the time.

The confusion matrix is also shown to showcase the classification and misclassification rate.

7. Model Comparison and Recommendation

In order to address the research problem of this study, it is important to select the right model to achieve optimal results. Model recommendation involves choosing a predictive model that best fits the specific characteristics of the data, the research problem, and desired outcome. The different models employed in this study included Classification Trees (decision tree and random forest), Logistic Regression (forward, backward and stepwise regression), Neural Networks, Naïve Bayes, and Support Vector Machine.

The selected model was based not only on the metrics but also on its efficiency, interpretability and scalability. The goal is to consider the model's performance in real-world scenarios, in this case Early Learning and Child Care. The carefully selected model helped provide deeper insights and drive better decision-making.

7.1. Model Selection

A total of 24 models were evaluated to identify the best performing one. Based on the results (found in table below), the following analysis was conducted.

Full Logistic Regression model (15 variables): The best model based on accuracy, recall/sensitivity and precision is the full logistic regression model with 15 variables. It had the highest accuracy of .72, which implies that the model provides 72% accurate results.

Model Name	ROC AUC	Accuracy	Recall	Precision	F1 Score
Full Logistic Regression	0.77	0.72	0.75	0.71	0.73

Table 21 Metrics for Full Logistic Regression

Full Logistic Regression model (9 variables): The second best model is the full regression with 9 variables from the feature selection of Random Forest model. It had an accuracy of 0.71, which implies that the model provides 71% accurate results.

Model Name	ROC AUC	Accuracy	Recall	Precision	F1 Score
Full Logistic Regression RF	0.75	0.71	0.75	0.70	0.72

Table 22 Metrics for Full Logistic Regression RF

If the study used the model performance metrics such as the accuracy score, it is obvious that the Full Logistic Regression model with 15 variables is the best model. However, the study chose the model that had a high enough accuracy, but with the best computational efficiency. Since the Full Logistic Regression model based on Random Forest feature selection has less variables (9), the study chose this model. A model with less variables has a computational efficiency because of the shorter training time needed. Shokrzad (2023) stressed that computational efficiency directly influences the operational cost of running the model (para. 1). Based on this combined model performance and efficiency, the study chose the Full Logistic Regression model based on Random Forest. It came in the second best model but had the best computational efficiency. A summary of the model comparison is found in the table below.

No.	Model Name	Accuracy	Recall	Precision	F1 Score	ROC AUC
1	Full Logistic Regression	0.72	0.75	0.71	0.73	0.77
2	Full Logistic Regression RF	0.71	0.75	0.70	0.72	0.75
3	Five-Depth Decision Tree	0.72	0.74	0.71	0.72	0.75
4	Forward Logistic Regression	0.72	0.71	0.73	0.72	0.74
5	SVM RF	0.71	0.77	0.69	0.73	0.75
6	Forward Logistic Regression	0.71	0.77	0.69	0.72	0.74
7	SVM	0.71	0.76	0.70	0.73	0.75
8	Backward Logistic Regression	0.71	0.75	0.70	0.72	0.76
9	Stepwise Logistic Regression	0.71	0.75	0.70	0.72	0.73
10	GridSearch Cross Validation	0.71	0.75	0.69	0.72	0.75
11	Backward Logistic Regression RF	0.71	0.75	0.69	0.72	0.75
12	Stepwise Logistic Regression RF	0.7	0.78	0.67	0.72	0.73
13	Five-Depth Decision Tree RF	0.7	0.76	0.68	0.72	0.74
14	Naïve Bayes	0.7	0.74	0.69	0.71	0.75
15	One-Depth Decision Tree	0.69	0.77	0.67	0.72	0.69
16	GridSearch Cross Validation RF	0.69	0.77	0.67	0.72	0.73
17	One-Depth Decision Tree RF	0.69	0.77	0.67	0.72	0.69
18	Naïve Bayes RF	0.69	0.76	0.67	0.71	0.74
19	Random Forest	0.69	0.70	0.69	0.70	0.74
20	Neural Network	0.69	0.70	0.69	0.69	0.72
21	Random Forest RF	0.68	0.69	0.69	0.69	0.73
22	Neural Network RF	0.67	0.68	0.68	0.68	0.71
23	Full Decision Tree RF	0.64	0.57	0.67	0.61	0.61
24	Full Decision Tree	0.63	0.58	0.65	0.61	0.61

Table 23 Summary of Metrics of all Models

7.2. Best Model

Full Logistic Regression with 9 variables from Random Forest is the best model in this study.

Model Name	Accuracy	Recall	Precision	F1 Score	ROC AUC
Full Logistic Regression RF	0.71	0.75	0.70	0.72	0.75

Table 24 Best Model Result

The confusion matrix below shows the recall and precision (based on validation data). Full Logistic Regression shows that the model identifies the use of child care 75% (recall) of the time, while the precision tell us when the model predicts the use of child care, the model is right 70% of the time.

Confusion Matrix (Accuracy 0.7083)

		Prediction	
	Actual	did not use CC	use CC
did not use CC		3385	1700
use CC		1293	3884

Figure 42 Confusion Matrix

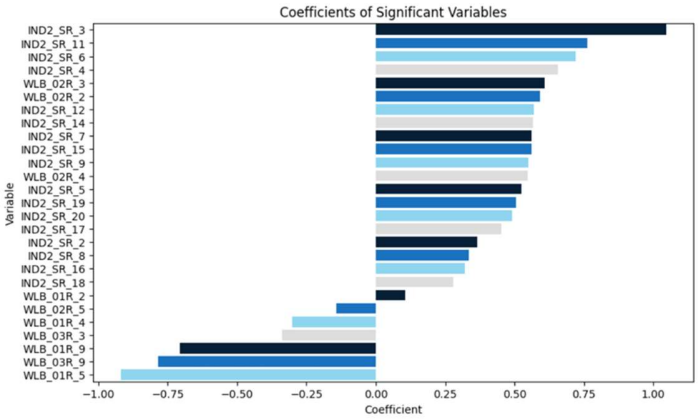

Figure 43 Coefficients for Significant Variables

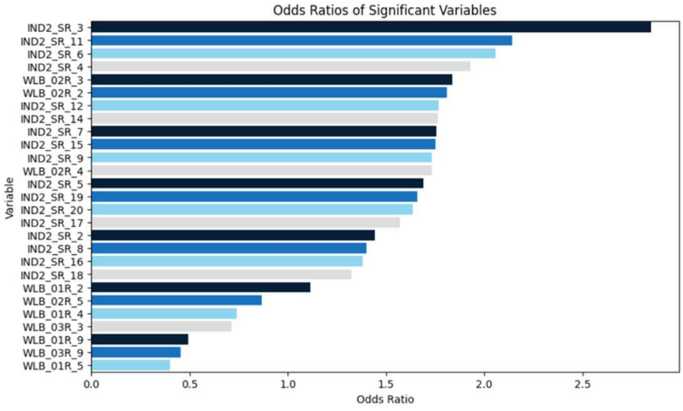

Figure 44 Odds Ratio of Significant Variables

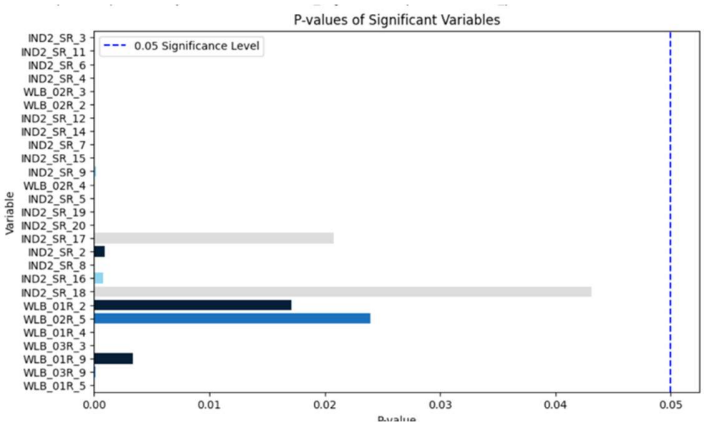

Figure 46 P-values for Significant Variables

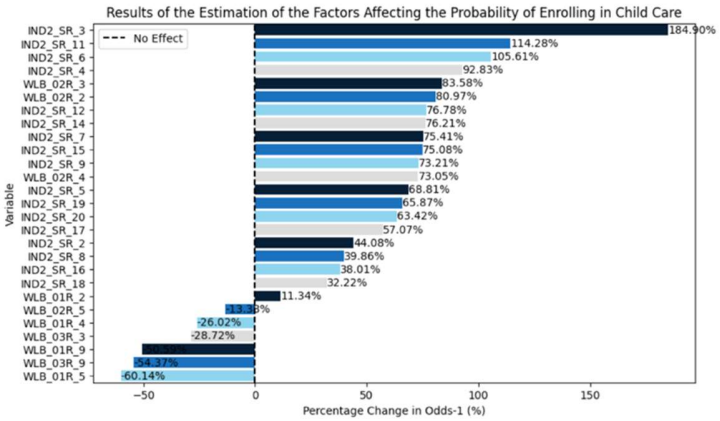

Figure 45 Result of Estimation

The table below shows the odds-ratio interpretation for all significant variables in the best model:

Variables	(Odds – 1)	Interpretation
WLB_02R_3	83.58%	Spouses who work in **info and cultural industries**, the probability for availing child services goes up by 73%
WLB_01R_5	-60.14%	Spouses who work in **health care and social assistance**, the probability for availing child services goes up by 38%
IND2_SR_12	76.78%	Parents who **RARELY** face difficulty fulfilling work responsibilities, the probability of availing child care goes up by 73%
IND2_SR_4	92.83%	Spouses who work in **transpo and warehousing**, the probability for availing child services goes up by 40%
WLB_02R_4	73.05%	Parents who **OFTEN** face difficulty to fulfill family responsibilities, the probability for availing child care increases 11%
WLB_02R_2	80.97%	Parents who **NEVER** face difficulty fulfilling work responsibilities, the probability of availing child care goes down by 13%
IND2_SR_20	63.42%	Spouses who work in **public admin**, the probability for availing child services goes up by 63%
IND2_SR_5	68.81%	Spouses who work in **retail**, the probability for availing child services goes up by 75%
IND2_SR_7	75.41%	Spouses who work in **manufacturing**, the probability for availing child services goes up by 69%
IND2_SR_6	105.61%	Spouses who work in **mining, quarrying and oil**, the probability for availing child services goes up by 44%
WLB_01R_4	-26.02%	Parents who **NEVER** face difficulty fulfilling family res due to work, the probability of availing child care goes down by 60%
IND2_SR_3	184.90%	Not stated, the probability for availing child care decreases 51%
IND2_SR_15	75.08%	Parents who **SOMETIMES** face difficulty fulfilling work responsibilities, the probability of availing child care goes up by 84%
WLB_03R_3	-28.72%	Spouses who work in **construction**, the probability for availing child services goes up by 93%
IND2_SR_8	39.86%	Spouses who work in **wholesale**, the probability for availing child services goes up by 106% or goes up twice as much
IND2_SR_19	65.87%	Spouses who work in **admin and support, waste management**, the probability for availing child services goes up by 76%
IND2_SR_11	114.28%	Spouses who work in **arts, entertainment**, the probability for availing child services goes up by 57%

IND2_SR_14	76.21%	Spouses who work in **other services**, the probability for availing child services goes up by 66%
WLB_03R_9	-54.37%	Parents who **RARELY** face difficulty fulfilling family res due to work, the probability of availing child care goes down by 26%
IND2_SR_9	73.21%	Spouses who work in **professional, scientific and technical services**, the probability for availing child services goes up by 77%
IND2_SR_16	38.01%	Not stated, the probability for availing child care decreases 54%
IND2_SR_2	44.08%	Parents who are neither satisfied nor dissatisfied with work life balance, the probability of availing child care goes down by 29%
WLB_01R_9	-50.59%	Spouses who work in **utilities**, the probability for availing child services goes up by 185% or goes up twice as much
WLB_01R_2	11.34%	Spouses who work in **education** , the probability for availing child services goes up by 75%
IND2_SR_17	57.07%	Spouses who work in **real estate, rental and leasing**, the probability for availing child services goes up by 114% or goes up twice as much
WLB_02R_5	-13.38%	Parents who **OFTEN** face difficulty fulfilling work responsibilities, the probability of availing child care goes up by 81%
IND2_SR_18	32.22%	Spouses who work in **accommodation and food services**, the probability for availing child services goes up by 32%

Table 25 Odds-Ratio Interpretation

7.3. Model Theory

To deepen the analysis of machine learning models, it is essential to incorporate the model theory. Since the study utilized the Logistic Regression and it turned out to be the best model, the study looked into the binary classification theory. The binary classification theory includes axioms that define rules for classifying inputs (whether 1 or 0).

- **Model Assumptions and Limitations**

Logistic Regression models require certain assumptions. Harris (2021) discussed several assumptions of logistic models. First, the outcome variable should be binary. In this study, the outcome or target variable is whether or not the parent had their child attend child care or not. The outcome variable is the dependent variable and should have only two possible outcomes: 1 (attended child care) and 0 (did not attend). Second, the features or independent variables in the model should be independent. This means that the outcome of one observation should not influence the outcome of the other. Third, there should be a linear relationship between the logit (log-odds) of the outcome and each feature variable. Fourth, there should be no perfect multicollinearity. This means that the feature variables should be perfectly correlated with each

other. Perfect multicollinearity can lead to problems in estimating the model coefficients. These assumptions were used to carry out the analysis of the results of the model.

One of the main limitations of the Logistic Regression model is that it assumes a linear relationship between the feature variables and the logit of the outcome. It the relationship turns out to be non-linear, the model may perform poorly.

7.4. Model Sensitivity to Key Drivers

Model sensitivity to key drivers involves understanding who changes in the feature variables affect the predictions of the model. First, coefficient magnitudes should be reviewed. This represent the change in the log-odds of the target variable for 1 unit change in the feature variable. Larger absolute values of coefficients indicate that a feature has a greater effect on the log-odds of the target variable. Second is feature importance. This measures how much each feature contributes to the model's predictions (Lundberg et al., 2017).

8. Conclusion and Recommendations

The exploration of child care utilization through predictive analytics has unveiled significant insights into the determinants that shape parents' and guardians' choices regarding child care arrangements. The study's findings highlight the complex interplay between economic, demographic, employment and parents' perception revealing which variables are most strongly associated with child care utilization as well as the direction of these associations. By understanding these relationships, the research provides a solid foundation for crafting targeted policy recommendations aimed at improving child care access in Canada.

8.1. Impacts on Business Problem

Dual-income dynamics

Based on the results of the Logistic Regression model, the labor participation of both parents significantly influences the decision to use child care services. Specifically, when both parents are employed, the likelihood of opting for child care increases markedly, regardless of the industry in which they are engaged. This finding underscores the critical role that dual-income households play in driving demand for child care.

Supporting this observation, Statistics Canada (2024) highlights a notable trend in labor force participation, particularly among women. In 2022, the labor force participation rate for women rose to 61.5% up from 58% in 1990. This increase reflects broader societal changes and economic factors that have led to greater opportunities for women. As more families experience dual-income dynamics, the demand for child care services is expected to rise correspondingly.

Specific industry drivers

Industries where spouse's employment is associated with a higher probability of availing child care include Utilities, Real Estate and Wholesale trade. In the utilities sector, structured and often demanding schedules can increase the need for reliable child care services. Similarly,

individuals working in real estate may face irregular hours and frequent client meetings, which can require additional child care support. In wholesale trade, the variability in work hours and demands of the job can often lead to greater need for dependable child care arrangements. Other industries that turned out to increase demand for child care are: construction, professional, scientific & technical, administrative support & waste management, retail trade, education, information and cultural services, manufacturing, public admin, arts, mining & quarrying, transportation, health care, accommodation and food services.

Balance between family and work responsibilities

Parents who experience minimal difficulty in fulfilling family responsibilities are significantly less likely to seek external child care. This is because such parents often have the flexibility to adjust their work arrangements to better accommodate family needs. For instance, they may choose to postpone their return to work, as they have the capacity to manage family responsibilities without additional support (Statistics Canada, 2022). Additionally, these parents might decide to work from home, allowing them to balance work and family duties more effectively. Furthermore, they may opt to work fewer hours to maintain a better work-life balance, which diminishes the demand for child care services.

8.2. Recommended Next Steps

Address growing demand especially in 'child care deserts'

To address the growing demand for child care, particularly in areas identified as 'child care deserts', it is crucial to enhance partnership between provincial and municipal governments in the public management and administration of child care services. Strengthening this collaboration will facilitate a more coordinated approach to increasing the supply of child care facilities and manpower. Key strategies include expanding the number of available child care centers and improving recruitment and retention of qualified staff. Additionally, increasing operational funding is essential to support these expansions and ensure high-quality care.

Investing in better compensation and working conditions for child care workers is vital for attracting and retaining skilled professionals, thereby improving the overall quality of care. Municipalities should take on a leading role in planning the locations and defining the characteristics of new child care facilities. This includes making critical decisions about expansion priorities and coordinating the acquisition of local permits, new licenses and reliable operating funding (Friendly et al., 2024). By focusing on these areas, municipalities can effectively address child care shortages and ensure that child care services are accessible and responsive to local needs.

Recognize the economic value of child care

Recognizing the economic value of unpaid child care is essential in addressing the broader implications of child care responsibilities. Unpaid care work, often performed by family members contributes significantly to the economy by support the paid workforce and enabling economic participation. Acknowledging the value can help inform policies and support systems that better recognize and reward the contributions of those providing unpaid care.

In light of this, it is important to raise awareness about the Employment Relations (Flexible Working) Act 2023, which introduces various flexible work arrangements (Janes, 2023). This Act provides employees with various options such as request flexible time, reduced hours or part-time work among others. These flexible options are crucial for supporting individual who balance work and caregiving responsibilities.

Incentives for child care providers and employers

To further support and promote effective child care solutions, it is crucial to implement positive incentives for employers. Recognizing and rewarding good workplace practices can motivate organizations to adopt and maintain high standards in supporting employees with child care needs. This can be achieved through awards and public recognition programs that highlight companies with exemplary workplace practices.

Capacity-building initiatives can play a significant role in promoting the sharing and dissemination of best practices. Additionally, offering tax incentives or grants to child care

providers who cater to employees in high-demand industries can encourage the expansion of services in areas with pressing needs. Finally, employers in sectors such as Utilities, Real Estate and Wholesale Trade should be encouraged to provide on-site or subsidized child care services. By doing so, these industries can enhance employee satisfaction and retention while addressing the child care needs of their workforce.

Incorporating these strategies can create a more supportive and effective framework for managing child care responsibilities, benefiting both employees and employers alike.

References

Friendly, M. Cleveland, G., Colley, S., Vickerson, R., Ferns, C., Holt, C. (2024). The municipal role in child care. Institute on Municipal Finance and Governance. https://imfg.org/wp-content/uploads/2024/05/wdwpaper_no8_child_care_may_2_2024.pdf

Harris, J. (2021). Primer on binary logistic regression. *Fam Med Community Health*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8710907/

Janes, S. (2023, August 1). Should Canada implement flexible working legistlation for all workpalces? *Benefits Canada.* https://www.benefitscanada.com/news/bencan/should-canada-implement-flexible-working-legislation-for-all-workplaces/

Lundberg, S. and Lee, S. (2021). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Mcdonald, D. (2023, May 29). Alarming statistics highlight child care accessibility crisis in Canada. *TroyMedia*. https://troymedia.com/lifestyle/alarming-statistics-highlight-child-care-accessibility-crisis-in-canada/

Shokrzad, R. (2023, November 29). The art of computation: time and memory optimization in ML. *Medium*. https://medium.com/@reza.shokrzad/the-art-of-computation-time-and-memory-optimization-in-ml-8848cbd5748b

Statistics Canada. (2022). Statistics Canada data strategy framework. https://www.statcan.gc.ca/en/about/datastrategy#a4

Statistics Canada. (2024). Canadian Survey on Early Learning and Child Care: Detailed Information for 2023. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5371

Statistics Canada. (2024). Canadian Survey on Early Learning and Child Care (CSELCC). https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5371&dis=1

Appendix

Sector (IND2_PR & IND2_SR)

Value	Sector	NAICS Code
1	Agriculture, forestry, fishing and hunting	11
2	Mining, quarrying, and oil and gas extraction	21
3	Utilities	22
4	Construction	23
5	Manufacturing	31-33
6	Wholesale trade	41
7	Retail trade	44-45
8	Transportation and warehousing	48-49
9	Information and cultural industries	51
10	Finance and insurance	52
11	Real estate and rental and leasing	53
12	Professional, scientific and technical services	54
13	Management of companies and enterprises	55
14	Administrative and support, waste management and remediation services	56
15	Educational services	61
16	Health care and social assistance	62
17	Arts, entertainment and recreation	71
18	Accommodation and food services	72
19	Other services (except public administration)	81
20	Public administration	91

Figure 47 Data Dictionary for Industry

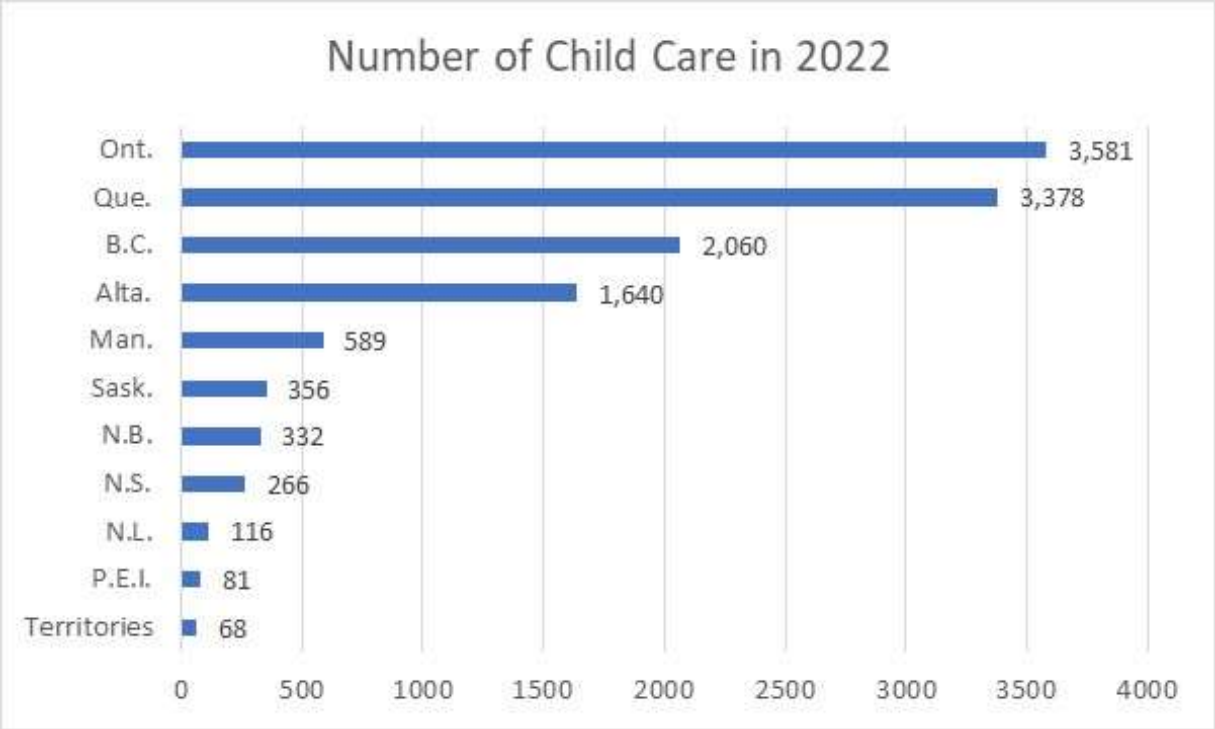

Figure 48 Number of Child Care Centers in Canada

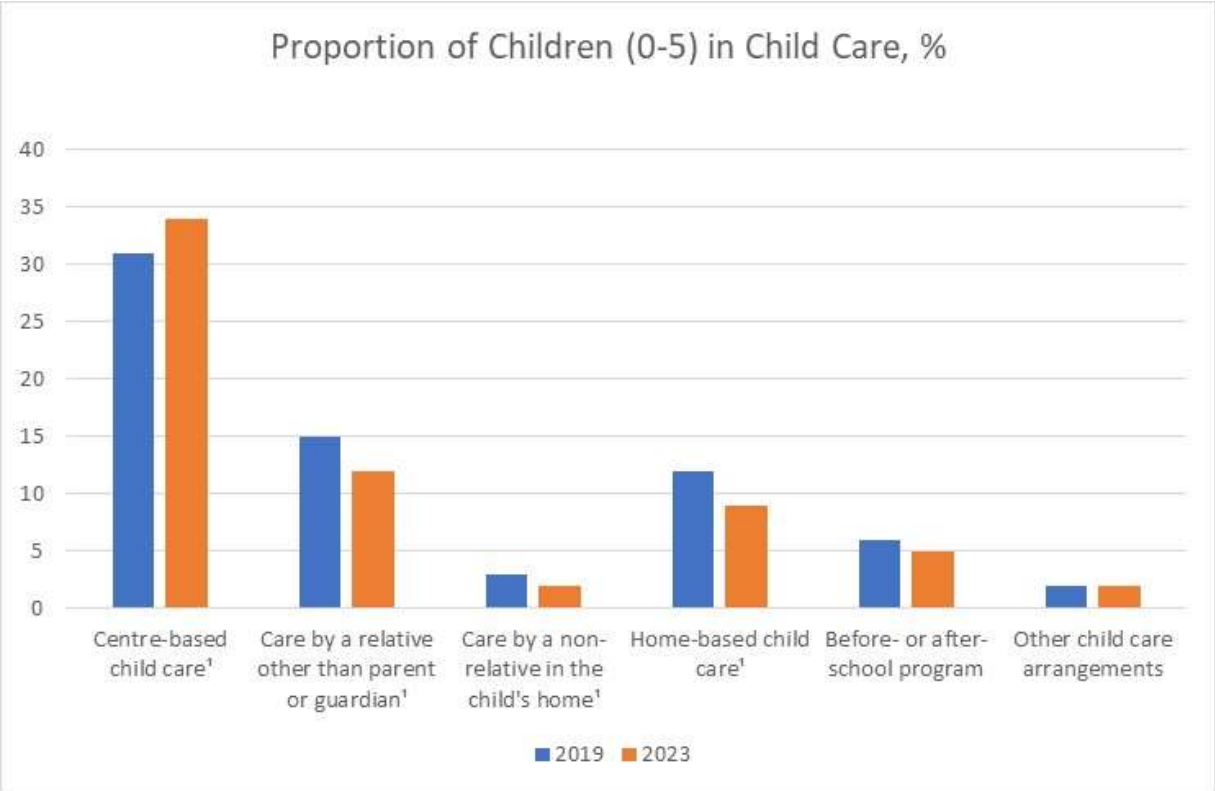

Figure 49 Types of Child Care 2019 vs 2023