# STAA57 Group Project

**What are the dominant research areas in Ontario's automotive sector, and how do institutions specialize in different fields**

Minh Tran - 1006804914, Swajeet Jadhav - 1009888276

2025-03-15

# I. Introduction

## Research Goal

With the growing popularity of electric vehicles (EVs) and the rapid advancements in artificial intelligence (AI), particularly in the development of autonomous vehicles like Tesla, it is crucial for automotive research facilities in Ontario to focus their efforts and resources on the right areas. This will ensure a significant contribution to the academia world while assisting the government in implementing effective policies and educational institutions better educate, prepare the next generation of labors.

This report aims to examine the automotive research areas prioritizing by Ontario's research facilities and identify specialized facilities in trending fields. It will also explore which areas are attracting the most researchers and which institutions are specializing in particular research domains.

## Dataset description

### Data source

Our data set was collected by researchers from the Ministry of Economic Development, Job creation and Trade and was published on the Government of Ontario's public database in 2018. A supporting description table of the research tag will be provided in the Appendix for better understanding of the main data set values.

### Data set Variables

```
## Rows: 545
## Columns: 37
## $ Institution                  <chr> "Brock University", "Carleton Univers~
## $ Researcher.Name              <chr> "Ahmed, Syed Ejaz", "Ahmadi, Mojtaba"~
## $ Associated.Facilities        <chr> "Centre for Statistical Consulting", ~
## $ Research.Areas               <chr> "Traffic and road injury prevention."~
## $ Research.Chairs.Grant.Funding <chr> "", "", "", "", "Canada Research Chai~
## $ Tag.1                        <chr> "Injury Prevention", "Mechatronics", ~
## $ Tag.2                        <chr> "", "Control", "Autonomy and AI", "Ne~
## $ Tag.3                        <chr> "", "", "Sensors", "", "", "", "", ""~
## $ Tag.4                        <chr> "", "", "", "", "", "", "", "", "", "~
## $ Tag.5                        <chr> "", "", "", "", "", "", "", "", "", "~
## $ Alternative.Fuels            <chr> "", "", "", "", "", "", "", "", "", "~
## $ Autonomy.and.AI              <chr> "", "", "x", "", "", "", "x", "", "",~
## $ Batteries.and.Fuel.Cells     <chr> "", "", "", "", "", "", "", "", "", "~
## $ Biocomposites                <chr> "", "", "", "", "", "", "", "", "", "~
## $ Coatings.and.Corrosion       <chr> "", "", "", "", "", "", "", "", "", "~
## $ Connected.Vehicles           <chr> "", "", "", "", "", "", "", "", "", "~
## $ Control                      <chr> "", "x", "x", "", "", "", "", "", "",~
## $ Crashworthiness              <chr> "", "", "", "", "", "", "", "", "", "~
## $ Electronics                  <chr> "", "", "", "", "", "", "", "", "", "~
## $ Forming.and.Joining          <chr> "", "", "", "", "", "", "", "", "", "~
## $ High.Strength.Steel          <chr> "", "", "", "", "", "", "", "", "", "~
## $ Hybrid.and.Electric.Vehicles <chr> "", "", "", "", "", "", "", "", "", "~
## $ Industrial.Processes         <chr> "", "", "", "", "", "", "", "", "", "~
## $ Injury.Prevention            <chr> "x", "", "", "", "", "", "", "", "", ~
## $ Internal.Combustion.Engines  <chr> "", "", "", "", "", "", "", "", "", "~
## $ Lightweight.Metals           <chr> "", "", "", "", "", "", "", "", "", "~
## $ Mechatronics                 <chr> "", "x", "", "", "", "", "", "", "", ~
## $ Nanotechnology               <chr> "", "", "", "", "", "", "", "", "", "~
## $ Networks.and.Security        <chr> "", "", "", "x", "", "x", "", "x", "x~
```

```
## $ Noise..Vibration.and.Harshness    <chr> "Err:507", "Err:507", "Err:507", "Err~
## $ Polymers.and.Composite.Materials <chr> "", "", "", "", "", "", "", "", "", "~
## $ Powertrain                       <chr> "", "", "", "", "", "", "", "", "", "~
## $ Sensors                          <chr> "", "", "x", "", "x", "", "x", "", ""~
## $ Software                         <chr> "", "", "", "x", "", "", "", "", "", ~
## $ Stress.and.Fracture              <chr> "", "", "", "", "", "", "", "", "", "~
## $ Transportation.and.Charging      <chr> "", "", "", "", "", "", "", "", "", "~
## $ Vehicle.Design                   <chr> "", "", "", "", "", "", "", "", "", "~
```

Based on a summary of the data set, the `Noise..Viration.and.Harshness` variable was corrupted during the process of uploading or collecting the data and therefore will not be used in our report

Some of the main categorical variables that can help categorizing the data include

- **Researcher.Name**

- **Institution**: Name of university or research center that the researcher works at

- **Research.Areas**: Researcher general field of research

- **Tag**: Researcher's specialized field of research

In addition, the data set also contains research fields as variable (e.g. Alternative Fuels, Autonomy and AI, Vehicle Design) to indicate whether or not a researcher work fall into that field of research

## II. Data Overview Analysis (Swaaa)

### *Descriptive Statistic*

- `Firstly lets try and change our data into some numeric values that can help us quantify and hence`

- Table 1: Summary statistics of research areas (count, frequency)

    - Identify the most common research fields

Table 1: Summary statistics of research areas

| Tag.1 | Count | Frequency |
|---|---|---|
| Networks and security | 67 | 0.1229358 |
| Autonomy and AI | 45 | 0.0825688 |
| Transportation and charging | 38 | 0.0697248 |
| Industrial processes | 34 | 0.0623853 |
| Software | 32 | 0.0587156 |
| Lightweight metals | 28 | 0.0513761 |
| Polymers and composite materials | 28 | 0.0513761 |
| Batteries and fuel cells | 26 | 0.0477064 |
| Connected vehicles | 23 | 0.0422018 |
| Forming and joining | 21 | 0.0385321 |
| Vehicle design | 21 | 0.0385321 |
| Alternative fuels | 18 | 0.0330275 |
| Biocomposites | 18 | 0.0330275 |
| Electronics | 16 | 0.0293578 |
| Nanotechnology | 16 | 0.0293578 |
| Control | 14 | 0.0256881 |
| Sensors | 14 | 0.0256881 |
| Coatings and corrosion | 13 | 0.0238532 |
| Internal combustion engines | 13 | 0.0238532 |
| Injury Prevention | 12 | 0.0220183 |

| Tag.1 | Count | Frequency |
|---|---|---|
| Hybrid and electric vehicles | 9 | 0.0165138 |
| Mechatronics | 7 | 0.0128440 |
| Noise, vibration and harshness | 7 | 0.0128440 |
| Powertrain | 6 | 0.0110092 |
| Stress and fracture | 6 | 0.0110092 |
| High strength steel | 5 | 0.0091743 |
| Crashworthiness | 4 | 0.0073394 |
| Other | 4 | 0.0073394 |

The above table provides summary statistics of research areas, displaying the number of researchers working in each field ("Count") and their relative proportions ("Frequency").

- Table 2: Top 5 institutions with the most researchers in the top research field
  - Identify the most dominant institution in a field

Table 2: Top 5 Institutions in Networks and security

| Institution | Reseacher_count |
|---|---|
| Carleton University | 33 |
| University of Waterloo | 15 |
| Ryerson University | 6 |
| University of Ontario Institute of Technology (UOIT) | 3 |
| University of Ottawa | 3 |

The table above will highlight the top five institutions with the most researchers working in Networks and Security.

### Graph & Visualizations

Graphs are a really important visual resprestation that helps us visaulize how the datset is distributed it allows us to analyse the variety of variables found in the data and how it relates to other variables across.

- Bar chart: Research distribution across institution

## Research Distribution Across Institutions
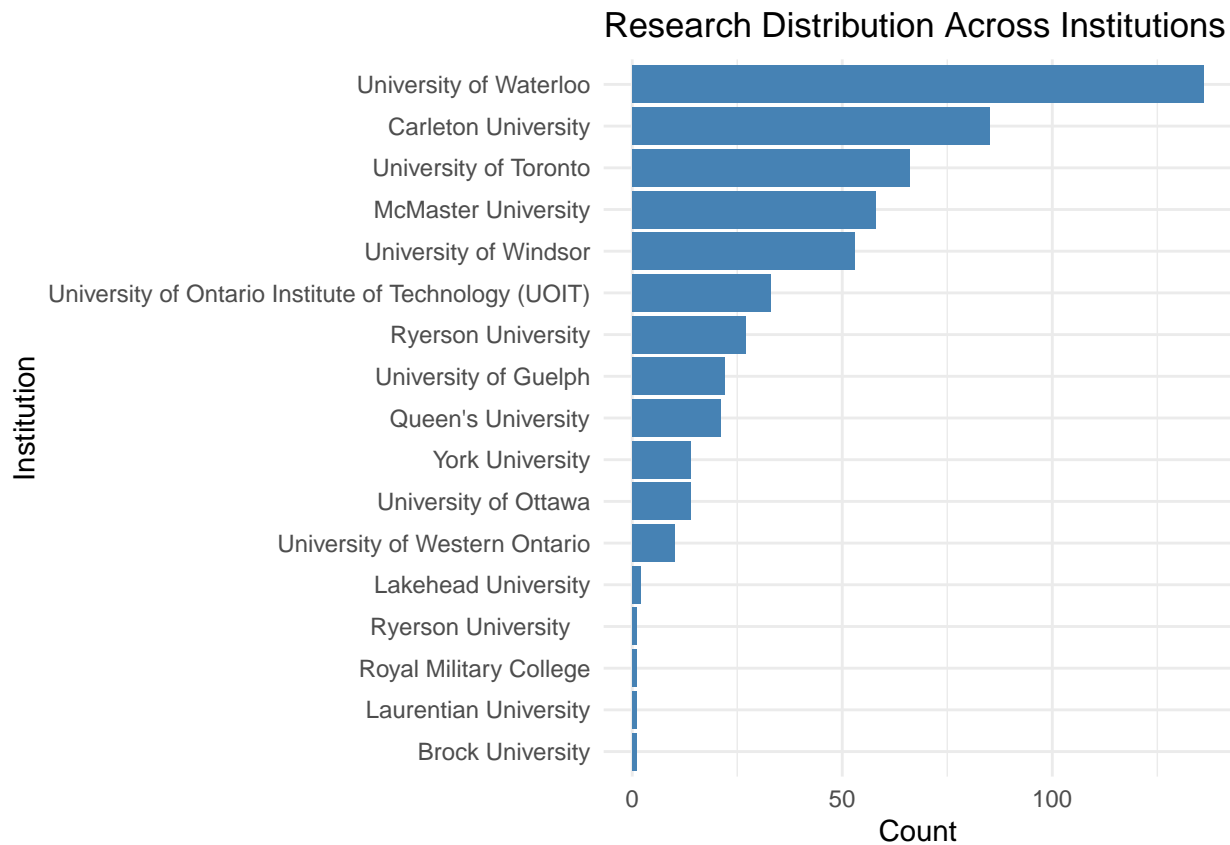
- The title "Research Distribution Across Institutions" clearly indicates the purpose of the graph.
- The x-axis (Institutions) and y-axis (Counts of Researchers) are labeled appropriately.
- Pie chart: Proportion of of different research fields

# Research Distribution in top Research Fields
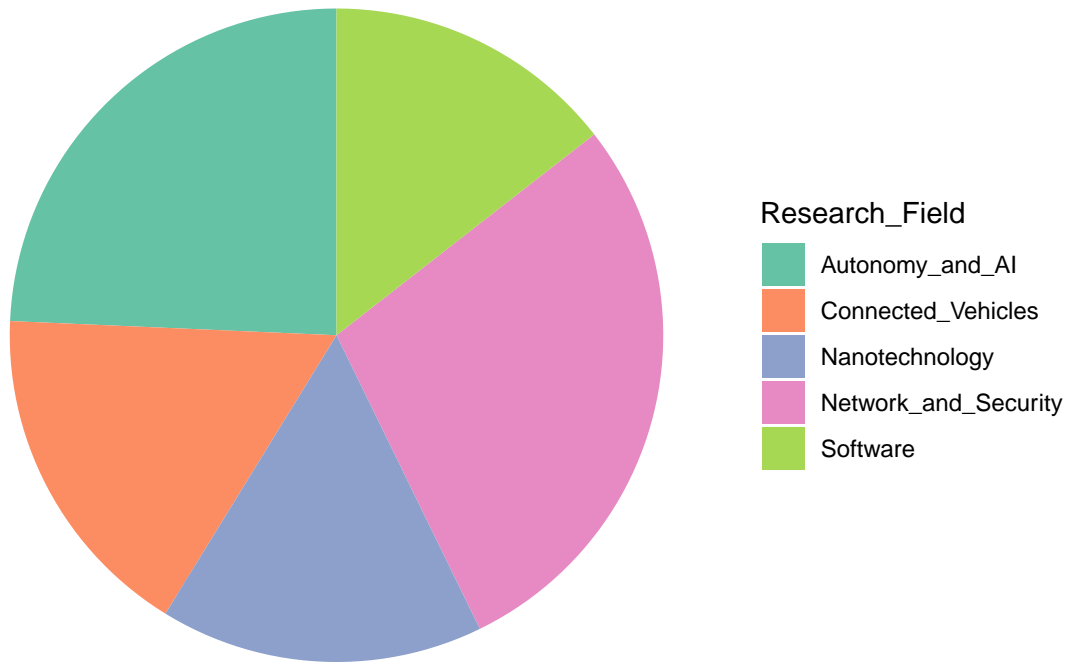


- Each slice represents the proportion of researchers working in that field, based on data aggregated from multiple institutions.

- History, frequency of research tag

## Frequency of Research Tag 1
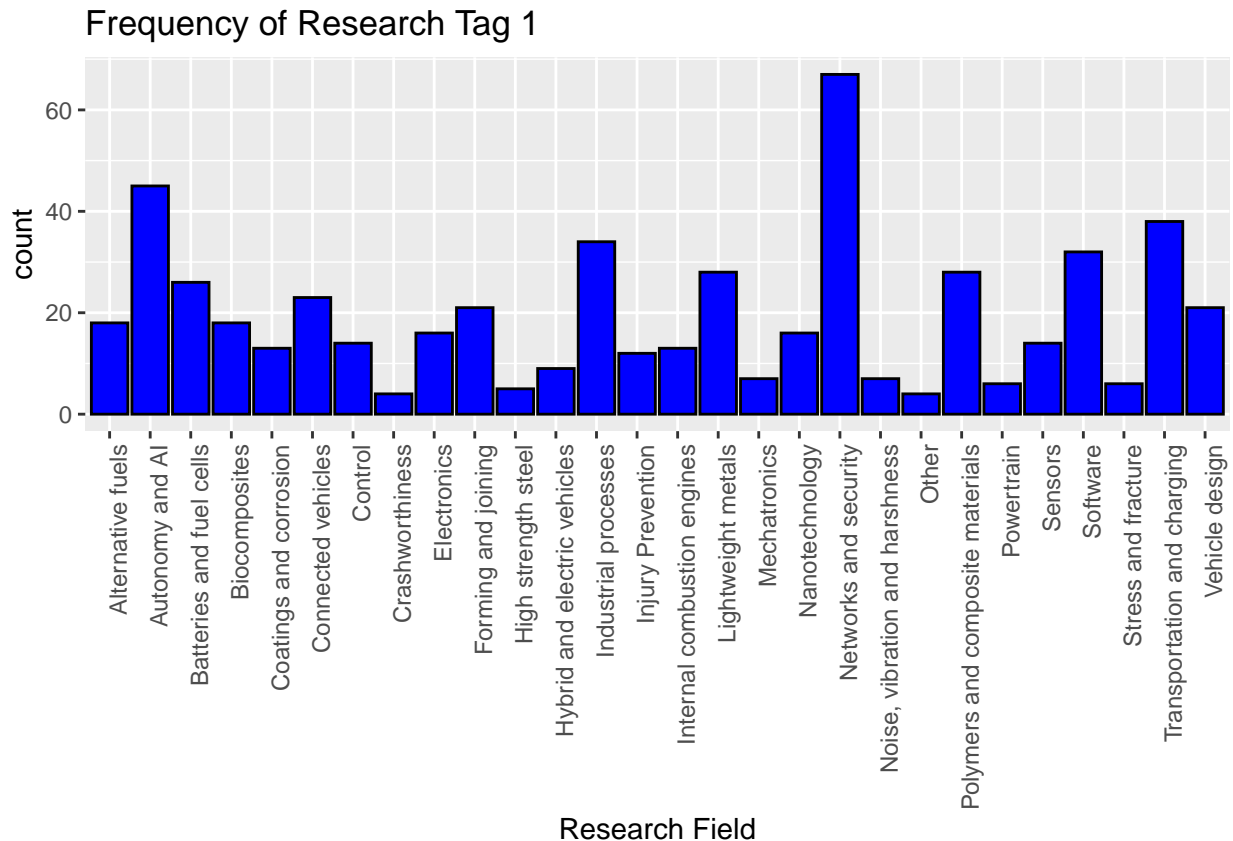


The graph is a bar chart showing the frequency distribution of different research fields based on the Tag.1 variable. - The graph likely has bars of varying heights, showing which research fields are more popular.

# III. Statistical Analysis (rayaanxsyed)

We could pick an area we think that will be most trending (AI for example) and test to see if that is true

- Confidence interval for the average number of researcher of all data set institutions

```
## [1] 16.34981 47.76783
## attr(,"conf.level")
## [1] 0.9
```

- Let's choose the most trending research topic (Network and security) and do a confidence interval for the average number of researchers in the field.

```
## [1]  0.3903509 10.4331786
## attr(,"conf.level")
## [1] 0.95
```

Hypothesis Testing For the topic (Network and security) and do a hypothesis test for the topic.

```
##
##  One Sample t-test
##
## data:  network_security_count
## t = 0.17384, df = 16, p-value = 0.8642
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##   0.3903509 10.4331786
```

```
## sample estimates:
## mean of x
##  5.411765
```

- Bootstrapping
- Bootstrapping is a re sampling technique used to estimate statistics on a dataset by repeatedly sampling with replacement. It helps assess the variability of an estimator (e.g., mean, standard deviation) and construct confidence intervals without relying on strong parametric assumptions. Do bootstrapping on network and security.

```
##      2.5%     97.5%
##  1.645588 10.530882

##
##  One Sample t-test
##
## data:  network_security_count
## t = 2.2847, df = 16, p-value = 0.03632
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   0.3903509 10.4331786
## sample estimates:
## mean of x
##  5.411765
```

# IV. Predictive modeling (Regression) (Minh)

**1. Logistic Regression Model**

*1.1. Model explanation*

The allocation of research funding plays a pivotal role in identifying key research fields that attract the attention of funding bodies such as Canada Research Chair Program and other major investors. These funding trend are essential indicators of the areas prioritized by the Canadian government and industry stakeholders in the automotive sector. Understanding these trends can provide insight into the direction of research investment, thereby influencing strategic decision in academia and industry alike.

To analyze the factors influencing research funding, we employed a logistic regression model with the primary dependent variable being **is_Funded**. This variable take a value of 1 if the research has secured funding from the Canada Research Chair program or other similar grants, and 0 otherwise. The independent variables include various research's primary fields as categorized in **Tag.1** column of the data set.

*1.2. Result and Key Findings*

**Table : Regression Model Result Summary**

```
##
## Call:
## glm(formula = is_Funded ~ ., family = binomial, data = reg_data)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -0.18232    0.60553  -0.301   0.7633
## is_Alternative.Fuels          1.43508    0.82951   1.730   0.0836 .
## is_Autonomy.and.AI           -0.04082    0.67577  -0.060   0.9518
## is_Batteries.and.Fuel.Cells   0.65233    0.72744   0.897   0.3699
## is_Biocomposites              0.40547    0.76920   0.527   0.5981
## is_Coatings.and.Corrosion     0.02817    0.82231   0.034   0.9727
```

```
## is_Connected.Vehicles                    0.81093     0.74722    1.085    0.2778
## is_Control                               -1.11696    0.88933   -1.256    0.2091
## is_Crashworthiness                        1.28093    1.30384    0.982    0.3259
## is_Electronics                            0.43364    0.78780    0.550    0.5820
## is_Forming.and.Joining                    1.62924    0.82188    1.982    0.0474 *
## is_High.Strength.Steel                   -0.22314    1.09545   -0.204    0.8386
## is_Hybrid.and.Electric.Vehicles          -0.04082    0.90370   -0.045    0.9640
## is_Industrial.Processes                   0.30010    0.69622    0.431    0.6664
## is_Injury.Prevention                      0.51879    0.84233    0.616    0.5380
## is_Internal.Combustion.Engines            1.38629    0.89443    1.550    0.1212
## is_Lightweight.Metals                     0.77011    0.72265    1.066    0.2866
## is_Mechatronics                          -0.10536    0.97468   -0.108    0.9139
## is_Nanotechnology                         0.69315    0.79582    0.871    0.3838
## is_Networks.and.Security                 -0.53330    0.65905   -0.809    0.4184
## is_Polymers.and.Composite.Materials       0.61764    0.71861    0.859    0.3901
## is_Powertrain                             0.18232    1.01653    0.179    0.8577
## is_Sensors                               -0.40547    0.82327   -0.493    0.6224
## is_Software                              -0.46431    0.71077   -0.653    0.5136
## is_Stress.and.Fracture                    0.87547    1.05672    0.828    0.4074
## is_Transportation.and.Charging           -0.35667    0.69265   -0.515    0.6066
## is_Vehicle.Design                         0.47000    0.74907    0.627    0.5304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 755.31  on 544  degrees of freedom
## Residual deviance: 704.64  on 518  degrees of freedom
## AIC: 758.64
##
## Number of Fisher Scoring iterations: 4
```

**Table : Regression Model Coefficient**

|                                     | coefficient | odd |
|-------------------------------------|-------------|------|
| is_Forming.and.Joining              | 1.6292405   | 5.10 |
| is_Alternative.Fuels                | 1.4350845   | 4.20 |
| is_Internal.Combustion.Engines      | 1.3862944   | 4.00 |
| is_Crashworthiness                  | 1.2809338   | 3.60 |
| is_Stress.and.Fracture              | 0.8754687   | 2.40 |
| is_Connected.Vehicles               | 0.8109302   | 2.25 |
| is_Lightweight.Metals               | 0.7701082   | 2.16 |
| is_Nanotechnology                   | 0.6931472   | 2.00 |
| is_Batteries.and.Fuel.Cells         | 0.6523252   | 1.92 |
| is_Polymers.and.Composite.Materials | 0.6176396   | 1.86 |
| is_Injury.Prevention                | 0.5187938   | 1.69 |
| is_Vehicle.Design                   | 0.4700036   | 1.61 |
| is_Electronics                      | 0.4336360   | 1.55 |
| is_Biocomposites                    | 0.4054651   | 1.50 |
| is_Industrial.Processes             | 0.3001046   | 1.35 |
| is_Powertrain                       | 0.1823216   | 1.20 |
| is_Coatings.and.Corrosion           | 0.0281709   | 1.03 |
| is_Autonomy.and.AI                  | -0.0408220  | 0.96 |
| is_Hybrid.and.Electric.Vehicles     | -0.0408220  | 0.96 |

|  | coefficient | odd |
|---|---|---|
| is_Mechatronics | -0.1053605 | 0.90 |
| (Intercept) | -0.1823216 | 0.84 |
| is_High.Strength.Steel | -0.2231436 | 0.80 |
| is_Transportation.and.Charging | -0.3566749 | 0.71 |
| is_Sensors | -0.4054651 | 0.67 |
| is_Software | -0.4643056 | 0.63 |
| is_Networks.and.Security | -0.5332985 | 0.59 |
| is_Control | -1.1169614 | 0.33 |

A substantial proportions of the variables int he model exhibit high **p-value** of greater than 5% suggesting that most of the research fields are not statistically significant predictors of whether a research project will be funded. This is likely due to the limited number of observations available int he data set, which may have constrained the model's ability to detect more nuanced relationships. Nevertheless, the model provides valuable insights into general trends in funding allocation by the Canadian government and major investors.

Despite the high **p-value** for many variables, certain research fields stood out interns of their impact on the funding probabilities. Specifically, **Forming and Joining**, **Alternative Fuels**, and **Internal Combustion Engines** demonstrated relatively high log-odds ratios. These fields exhibited 4 to 5 times higher odds of receiving funding than those in other areas. In addition, **is_Forming.and.Joining** and **is_Alternative.Fuels** variables were both statistically significant with p-values of approximately 4% and 10% respectively which further support the findings above.

The significance and high odd of **Forming and Joining** research fields aligns with ongoing research efforts to improve manufacturing process which are crucial for the production of commercial vehicles. While the corresponding number for **Alternative Fuels** underscores the importance of environmentally sustainable technologies, which are increasingly emphasized in both governmental policy and industry innovation.

**2) Cross validation**

To further evaluate the robustness of the logistic regression model, we conducted a k-fold cross-validation with four folds.

The results are as follow:

## AUC score: 0.6137124 0.6056287 0.5370879 0.6214709

## Average AUC score: 0.5944749

The model predictive performance assessed using the AUC score. The resulting average AUC score was 0.609, which suggests a fair to somewhat weak performance. This moderate performance is likely attributed to the large number of non-significant variables in the model, reflecting the challenges posed by limited data. Nevertheless, the model remains useful for identifying general funding trends rather than providing highly accurate predictions for individual research projects. It serves its purpose of illustrating that dominant research areas that are attracting funding, which is the primary objective of the analysis.

# V. Summary

In conclusion. . .

The logistic regression analysis provides valuable insights into the research funding landscape within the Canadian automotive sector. The significant fields of **Forming and Joining** and **Alternative Fuels**, along with the notable importance of **Internal Combustion Engines**, underscore the ongoing focus on traditional automotive manufacturing processes and sustainable technologies. These finding aligns with the broader trends observed in the automotive industry, where there is increasing attention to environmentally friend alternatives to conventional fuels and more efficient manufacturing practices.

However, as the market of Autonomous Vehicles and Electric Vehicles grows, we believe that more resources should be reallocate to support researches in these emerging areas. Fields such as **Autonomy and AI**, **Batteries and Fuel Cells**, and **Hybrid and Electric Vehicles** are expected to play a pivotal role in shaping the future of automotive industry. # Appendix