# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | Auto insurance customer segmentation based on profile using Clustering |
| **Requestor** | Business School, Centennial College |
| **Date of Request** | 8th July, 2022 |
| **Target Quarter for Delivery** | 22nd August, 2022 |
| **Epic Link(s)** | Not Applicable (We do not have Agile group for this project*)* |
| **Business Impact** | Insurance companies’ rivalry has become more intense over the years, Customer retention and acquisition are critical for maintaining a competitive advantage in the industry. This project will help the automobile insurance firm price their auto insurance competitively, which involves having a strong model for consumers who are at danger of being involved in an accident.  It will further assist in determining the coverage that should be offered based on specific traits or commonalities discovered in distinct consumer groups or segments. |

This project aims to segment insurance customers based on variables such as age, gender, employment details, education level, number of dependents, and so on, and to identify the unique characteristics of each segment that can present a better model for underwriting an auto insurance policy that is more cost effective and reasonable than using postal codes.

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Customer service is an essential component of insurance administration. As a result, Insurance companies must determine the critical success criteria in offering auto-insurance policies for their client’s happiness. This project will segment insurance customers based on variables such as age, gender, employment details, education level, number of dependents, and so on, and identify the unique characteristics of each segment that can present a better model for underwriting an auto insurance policy that is more cost effective and reasonable than using postal codes. |

**The specific tasks:**

In this project, I will perform the under listed tasks:

1. Identify different groups of customers using different modeling techniques.
2. Present a segmentation model for the auto-insurance company.
3. Determine the most acceptable customer segmentation technique.
4. Identify the key characteristics of customers in each group.
5. Develop customized auto insurance packages.
6. Assess the risk level of each segment.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | According to Statista.com, the share of net premium written in Canada for vehicle insurance between 1990 and 2020 is 26.96 billion CAD, a figure that is always rising due to current inflation. Furthermore, the net claim incurred by vehicle insurance for the same period is 17.86 billion CAD, which is more than 65 percent of the premium received during the same period. This does not appear to be a profitable venture. Many vehicle insurance customers complain of excessive rates owing to the usage of postal code information to compute premiums in the middle of these high premiums and claims. There is a need for a fairer and more transparent car insurance industry to increase profitability and consumer happiness, and a more robust model to determine auto insurance premiums. |

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | By completing this project, insurance companies will have a better model that will help in underwriting auto insurance premiums other than relying on postal code which lumps very different customers together. Since this project aims at grouping only similar customers together, their unique characteristics will help determine which type of customer to classify as high risk, low risk and medium risk, thereby reducing the unfairness of determining premiums through postal codes. |

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | Assuming that there are no duplicates, the dataset used for this project contains 10,000 customer records from an auto insurance company with 19 variables including demographic information, past accident, credit score, number of speeding tickets, number of convictions on driving under the influence (DUI) and past claims. The customers will be divided into various groups. Each group will only have members who are very similar, and members of different groups would be extremely dissimilar.  We would use both supervised and unsupervised algorithms to make various splits or clusters. The first sets of splits will be done using Decision Trees  The second sets of splits will be done using K-means. We would compare these clusters produced from both algorithms to domain knowledge of risk levels in auto insurance companies in order to come up with a suitable modeling technique for customer risk levels. |

## 2.1 Other related questions and Assumptions:

* The datasets we are using is correct and meets all quality standards
* The property and casualty insurance sector is willing to adjust premium calculation methods as a result of the outcome of this project
* Our model accurately captures all categories of insurance customers.
* Available data is recent, and recommendation can be applied in coming years.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | * Accurate classification of each customer in the correct group * Individual customers in each group must have similar characteristics * Individual customers in a group must have dissimilar characteristics with individuals in other groups. |
|  |  |
|  |  |

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | * Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology. |

**Type of Analysis:** *K-means Clustering, PCA (Principal component analysis),*

The initial approach will be to explore the dataset and apply the describe function to see descriptive statistics about the dataset. We will also use standardization technique if required and calculate the correlation between our variables before running our analysis.

**Methodology:** *Key questions from ‘Analytics objective’ will be tackled in ascending order as outlined in ‘5.0 Timelines and deliverable section’.*

We will begin by importing the data/python packages and then perform exploratory data analytics by plotting some graphs to see how our clients are distributed in terms of gender, income, age, and so on. The EDA results will assist us in determining outliers, most important variables, correlated variables, and then we will choose the value of K for our clustering by first using the 'apriori' or domain knowledge of three clusters (the three clusters are derived from the general knowledge that insurance customers are classified as high, medium, and low risk customers), and we will use this expected knowledge to inform our choice of cluster numbers.

Next, we'll take a more conventional technique to determining the number of clusters.We will first utilize the elbow approach, then the average Silhouette method, and finally the Calinski-Harabasz Score method to determine the appropriate number of clusters. We will also look for the WCSS (within cluster sum of squares) for the clusters formed by these statistical approaches.

These statistical approaches will instruct us on the best number of clusters to employ, and we will select the clusters with the lowest WCSS.

In the following phase, we will examine the similarities and dissimilarities that exist between objects in the same clusters. These commonalities and differences will guide us in labelling these clusters based on risk levels.

**Output:** The output will be a set of insights, observations and recommendations that will help the auto insurance industry with fair automobile insurance premium pricing.

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection: Insurance customer for Auto-Insurance company**

**Observation window: Not provided**

**Inclusions: Age, Gender, Driving Experience, Education, Income, Credit score, Vehicle ownership, vehicle year, married, children, Vehicle type, speeding violations, DUIs, past accidents, outcome.**

**Exclusions: Id, Postal code, Annual Mileage**

**Data Sources:** Sourced from Kaggle.com <https://www.kaggle.com/datasets/racholsan/customer-data?resource=download&select=customer-data.csv>

**Audience Level:** Students, professors, Insurance companies

**Variable Selection: PCA**

**Derived Variables: Not applicable**

**Assumptions and data limitations:**

Data Limitation:

* Time constraints
* Dataset has no date specifically year.

Assumptions

* The datasets we are using is correct and meets all quality standards
* The property and casualty insurance sector are willing to adjust premium calculation methods because of the outcome of this project
* Our model accurately captures all categories of insurance customers.
* Available data is recent, and recommendation can be applied in coming years.

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| *Clustering data of varying sizes* | *Low* |  | *Once analysis begins, we will generalize k-means. This approach allows us to satisfy some constraints.* |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Kick-off / Formal Request | *Introducing data and project objective to Advisor and program coordinator.* | *Provide dataset link.*  *Identify what the project sets to achieve.*  *Discuss how to improve project idea.* |  | *4th July, 2022* |
| 2. | Assessment / Triage | N/A | N/A |  |  |
| 3. | Prioritization | N/A | N/A |  |  |
| 4. | Data Exploration & Analysis   * Issues with duplicates * Issues with Spend data | Identifying variable and data types, treating missing values if applicable, removing redundant variables and visualization of variables through charts/graphs to present distribution of our variables. | *The EDA results will assist us in determining outliers, most important variables, correlated variables, and then we will choose the value of K for our clustering.* |  | *22nd July, 2022* |
| 5. | Story Board 1 | Visual presentation depicting our project. Frame key messages for our data story. | *Narrative storyboard framing research problem and plan to answer identified questions.* |  | *29th July, 2022* |
| 6. | QA Output |  |  |  |  |
| 7. | Internal team Presentation | Present project to classmates and professors. | *Provide slides that reinforces project objectives and present to classmates/professors* | *1* | *8th August, 2022* |
| 8. | Go/No Go | N/A | N/A | N/A |  |
| 9. | Story Board 2 | N/A | N/A | N/A |  |
| 10. | Pilot | N/A | N/A | N/A |  |
| 11. | Delivery & sign-off | Sign off on all project deliverables and close the project. | *Submit dataset, project codes and documents.* |  | *15th August, 2022* |