# Analytics Startup Plan

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

|  |  |
| --- | --- |
| **Project** | *Detection of Fraudulent Vehicle Insurance Claims* |
| **Requestor** | *Centennial College* |
| **Date of Request** | *July 11, 2022* |
| **Target Quarter for Delivery** | *August 18, 2022* |
| **Epic Link(s)** | *Not Applicable: As this Analysis doesn’t affect any agile group* |
| **Business Impact** | *Detecting fraudulent claims will help the business to prevent itself from paying for false claims. Once proven fraudulent, the insurance company may cancel the insurance. Moreover, other insurance companies may refuse to sell or charge higher prices (premiums) for selling insurance to people involved in false claims. This will aid in the expansion and effective functioning of the company.* |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

Insurance fraud has accompanied insurance since its inception and the frequency of insurance fraud incidents have been increased over time. Vehicle insurance fraud involves conspiring to make false or exaggerated claims involving property damage or personal injuries following an accident. Some common examples include staged accidents where fraudsters deliberately "arrange" for accidents to occur. Many Insurance companies have to face a decline in earnings, reserve deficiencies, rising loss costs, pricing difficulties, and other insurance expenses. This analysis will help in detecting fraudulent claims from the dataset provided.

**The specific ask:**

*Clearly articulate the specific task you will be conducting to help achieve the opportunity*

1. To develop a model that will predict fraudulent claims.
2. To identify the drivers that are driving fraud.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

Auto/vehicle insurance is the insurance for cars, trucks, motorcycles, and other road vehicles. Its principal role is to provide financial protection from risks like property damage and bodily injury resulting from liability arising from incidents in a vehicle. Property and Casualty insurance covers 51.1% share of total annual premiums received, and auto insurance is one type of property and casualty insurance product. In addition, the Auto Insurance industry as of 2022, is worth about $316 billion. The dataset contains information about the insurance policy as well as about the consumer. Also, the data set provides information about the person who was at fault. This dataset contains vehicle details (attributes, model, etc.) along with policy details (type, tenure, etc.). The target is to detect if a claim application is fraudulent or not.

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

The purpose of this analysis is to detect the fraudulent claims made from the dataset that will help in classifying the policyholders to whom the compensation should be made. Moreover, it is important to perform this analysis to increase the company’s earnings and reserving potential. Likewise, there is no need of wasting resources on assigning a specialist to analyze every single claim. However, if this analysis is not performed, then, the company have to face pricing difficulties and have to charge more even to those consumers who are not even submitting intimation for the claim.

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.*

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

The analytics Objective of this analysis is as follows: -

1. The primary objective is to find the best model with the highest accuracy rate for detection of the fraudulent claims.
2. The analysis will be delivering data visualizations.
3. The analysis will be taking into consideration the correlation factor for data reduction.
4. The analysis will deliver various predictive models such as decision tree, random forest, regression models.
5. The analysis will state the factors that will identify the detection of frauds.

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

The following are the dataset-related factors that may affect the analysis: -

1. The fraud claims (923) are significantly less common than legitimate insurance claims (14497). So, will have to take this into consideration while analyzing.
2. Age column contains some zeros, assuming that there was some other person driving the car so will have to replace with legitimate values.

The following are the assumptions that may affect the analysis:

1. The personal information about the policyholder provided in the dataset is legitimate.
2. The testimony provided by the witness of the accident in the dataset is assumed to be true.
3. The data is stationary through the observation window mentioned (1994-1996).
4. The data is the correct representation of the time frame.
5. “Year” column is assumed to be the year in which the accident occurred.
6. “Days\_Policy\_Accident” column is assumed to be the number of days between when the policy was purchased, and the accident occurred.
7. “Days\_Policy\_Claim” column is assumed to be the number of days that pass between the policy was purchased and the claim was filed.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  | The following are the Key Performance Indicators (KPI): -   1. Increase in earnings of the company. 2. Decrease in the premium charged by the insurance company. 3. Decline loss costs occurred due to fraudulent claims 4. Increase in reserves of the company. 5. The best model will be chosen on the basis of root mean square error (RMSE). 6. Increasing customer satisfaction as a proper model will be used to detect fraudsters and innocents will not be blamed. |
|  |  |

## 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:** Decision tree, Random Forest, Neural Networks, and Regression models

Initially, I will be creating decision trees or ADA fit to analyze the feature importance and are most significant. Moving further, I will also use other techniques to verify the findings.

**Methodology:** The key factors and assumptions from “Analytics objective” will be taken into consideration and the analysis will be performed as outlined in “ 5.0 Timelines and deliverable section’.

I will first start with exploratory data analysis (EDA) as the data is not much clear and have to fix it accordingly. FraudFound\_P is my target variable. Moving further, I will be employing feature engineering in order to transform data into features. Moreover, I will be building decision trees, random forests, neural networks, and regression models. At last, I will be selecting the best model on the basis of root mean square error among them to predict the results.

**Output:** The output of this analysis will be the classification and detection of fraud claims. Furthermore, the output will also provide an insightful conclusion of the analysis and strategic recommendations that will help the company to achieve its KPIs mentioned in 2.2

## 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:** Not applicable.

**Observation window:** Taking into consideration accidents occurred for the year 1994 to 1996.

**Inclusions:** Auto/Vehicle insurance claims data.

**Exclusions:** All the variables in the data set but excluding PolicyType column.

**Data Sources:** Kaggle: <https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection>

**Audience Level:** Not Applicable.

**Variable Selection:** Selecting FraudFound\_P as target variable.

**Derived Variables:** The following are the derived variables.

1. Bucketing “Age” into a new column.
2. Mean of the Vehicle price range will be considered.
3. Dummy variables will be derived.

**Assumptions and data limitations:** Assumptions about the data set:

1. “Year” column is assumed to be the year in which the accident occurred.
2. “Days\_Policy\_Accident” column is assumed to be the number of days between when the policy was purchased, and the accident occurred.
3. “Days\_Policy\_Claim” column is assumed to be the number of days that pass between the policy was purchased and the claim was filed.
4. Data is distributed as random sample.

## 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** |
| 1. *AgeofpolicyHolder column is the grouped class of Age column that we can see is not accurately done.* 2. *VechileCategory and BasePolicy are derived from the column PolicyType* | *Medium*  *Medium* | *1-2 days*  *1-2 days* | 1. *Once analysis begins, I can fix the range of the column AgeofpolicyHolder, accordingly.* 2. *Once started with the EDA, columns can be dropped to achieve data reduction or improve data quality.* |
| 1. *Delay in peer reviewing* | *Low* | * 1. *days* | 1. *Reduce the scope for visualization and quality of presentation.* |

## 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 | *Analysis plan discussion.* | *Laying a blueprint to follow* | *4* | *July 14, 2022* |
| 2. | Exploratory data analysis (EDA) | *Summarizing the main characteristics.* | *Data quality descriptive statics* | 6 | *July 18, 2022* |
| 3. | Data Analysis and preparation   * Issues with ambiguous data found. * Issues with wrongly derived columns. | 1. *Dealing with ambiguous zeros in various columns.* 2. *Dealing with existing columns, where exists wrongly derived values*. | *Cap and floor*  *Transform unusual data*  *Imputing the values* | 8 | *July 24, 2022* |
| 4. | Modeling | *Building statistical models and assumptions to detect fraud.* | *Building Decision tree, Random Forest, Neural Networks, and Regression models* | *5* | *August 1, 2022* |
| 5. | Report Preparation | *To begin with report preparation.* | *Preparing a formal report while defining problem statement, objectives, techniques and so on.* | *8* | *August 6, 2022* |
| 6. | Governance and Documentation | *Documenting the report while following a proper governance.* | *Story telling methods will be deployed with a proper framework* | *4* | *August 13, 2022* |
| 7. | Presentation | *Delivering the final presentation of the analysis performed.* | *Delivering the formal Presentation with full disclosure of qualitative and quantitative facts* | *1* | *August 17, 2022* |
| 8. | Populate project repository | *Finally, submitting the report for grading in the assignment folder.* | *The report in a pdf/word format will be submitted.* | *1* | *August 18, 2022* |