# Analytics Startup Plan

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

|  |  |
| --- | --- |
| **Project** | *Incorporating alternative consumer data in Credit Risk Modeling* |
| **Requestor** | *Centennial College* |
| **Date of Request** | *July 12, 2023* |
| **Target Quarter for Delivery** | *August 16, 2023* |
| **Epic Link(s)** | *Not Applicable* |
| **Business Impact** | *The success of this project aims to empower credit inclusion by using alternative data to predict the probability of default of underserved customers. Identifying key characteristics of default will help ensure that underserved customers who are capable of repayment will not be rejected.* |

## 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. |

Consumers take out loans for a variety of reasons such as paying for education, starting up a business, buying a car or a home, or big purchases of other products. Some people, even the wealthy ones, also prefer loans to save their available cash for future emergencies. Loans also play an integral part in a bank’s revenue model since most of its income is from the interest generated from loans. However, providing loans is subject to credit risk or the probability that the lender will not be paid on a loan and lose the money to a borrower. One way that lenders use to manage credit risk is by assessing the probability of default of a loan applicant also called default risk. Default risk is defined as the likelihood that a borrower won’t be able to make their required debt payments to a lender (Kagan, 2023).

For individual consumers, default risk can be measured based on their credit reports and credit scores. Credit bureaus gather and prepare credit reports from the information being sent by an individual’s previous and current lenders which shows their diligence in paying loans. Based on these reports, the credit score will then be calculated which will then affect the individual's future loan applications, rental applications, credit card applications, car leasing, and even interest rates. The assumption is that a consumer who has established a record of paying their bills on time is less likely to default in the future; therefore, lenders prefer a good credit score to minimize credit risk.

The dependency on credit scores for default risk assessment of most financial institutions creates a setback for credit-disadvantaged customers who accounts for more than 9 million Canadian consumers or 31% of the adult population (TransUnion, 2023). These are the credit unserved and underserved consumer segments who are struggling to access financial products and services due to no, or not enough, credit history. For instance, new immigrants in the country usually have a slimmer chance of getting approved for a car loan or if not, have higher interest rates since they have lower credit scores.

**The specific ask:**

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

This project aims to incorporate the use of alternative consumer data to predict loan repayment capabilities of the underserved population, ensuring that those who are capable will not be rejected. Understanding the overall financial capacity of underserved consumers will help lenders offer more credit to this segment, therefore generating new credit applications and increasing the customer base.

## 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? |

This project will build on the previous study on improving credit risk modelling with the collective use of alternative and traditional data in granting credit applications (FICO, 2022). The study defined traditional data as any information from credit bureaus, a credit application or a lender’s own files on an existing customer, which is the most used for credit scoring. On the other hand, alternative data refers to any data that is not directly related to a consumer’s credit behaviour. Some examples of alternative data are a) transaction data from credit or debit cards, b) Telecom / Utility / Rental data which are considered alternative because it's not existing in credit reports, c) Social Profile data, d) Clickstream data or their online behaviour on the website, and e) Audio and text data such as recorded customer service calls. This study from the FICO blog reports that combining the traditional and alternative data characteristics enabled them to produce a more powerful predictive model.

The dataset to be used is collated by Home Credit, an international non-bank financial institution which operates in 9 countries across Europe and Asia. Their mission is to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. They offer point-of-sales loans, cash loans and revolving loans to consumers with little or no credit history by using alternative data such as telco and transactional information to predict clients’ capability to fulfill their loans.

## 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.* |

Traditional data is commonly used in credit scoring and credit risk analysis of applicants; therefore, conducting this project will help reveal the key characteristics of predicting default from available alternative client information. Doing this project can help provide credit opportunities for the underserved population and in turn will increase the company’s market share by tapping on this credit-disadvantaged market segment.

## *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 main purpose of this project is to gain insights from the client’s traditional and alternative data to reveal the key characteristics that are important in predicting a credit default.

The second objective of this project is to create different predictive models and identify the best model that will accurately classify a credit-worthy customer from a potential defaulter.

Finally, this project will also aim to improve the performance of the machine learning model to ensure that our final model will provide the highest accuracy, as the cost of making a mistake in predicting a default can be very high for the company.

The project will seek to test the hypothesis: How the use of alternative data in credit risk modelling will help financial institutions determine credit-worthy customers even with little or no credit history in the country.

Key questions:

1. What key characteristics have the highest importance in determining a default?
2. What variables can accurately classify a defaulter and a non-defaulter?
3. How can we be sure that the model provides the optimal solution (minimizing the cost of error)?

## 2.1 Other related questions and Assumptions:

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

Assumptions:

1. Some variables in the dataset are not allowed to be used for credit decisions in some countries as it might be considered discrimination (i.e. age, gender). In this project, we will assume that all variables are free and acceptable to be used in modeling.
2. The dataset is gathered from the countries where Home Credit is operating so some variables might or might not be available in other countries. The techniques to be used and final model cannot be directly applied to a new country without rebuilding, retesting, and refitting the model.

## 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 cost of making an error in a default prediction can be very high due to the large amount of funds associated to each loan application; therefore, to evaluate the performance of the models, the following metrics will be used:

1. Confusion Matrix
2. Recall Score
3. Precision Score
4. Accuracy
5. F1 Score
6. ROC-AUC Score

## 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:** *Classification problem*

*Different types of classification models will be used to determine which variables are most significant in predicting if an applicant will default on a loan. Some of these techniques are decision trees, logistic regression, neural network, and ensemble models.*

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

1. **Data preparation**

This step includes downloading the datasets from the source and loading to Python in correct format (i.e. xlsx, csv).

1. **Data preprocessing / EDA**

Data preprocessing would start with assessing data structure and validating data quality like duplicates, missing values, skewness, and incorrect data types. This will be followed by data cleaning based on the identified issues in the data quality. Data transformation will be done to fix outliers and skewness on the data set. Data reduction includes removing unwanted variables that has no impact in modelling like variables with a single unique value and transaction number.

1. **Feature Engineering**

Creation of new features if needed.

1. **Modeling and Performance Evaluation**

Multiple machine learning algorithms will be created and will be compared based on the performance metrics mentioned in ‘2.2 Success metrics’. The best model will be chosen and deployed.

**Output:** *The output will be a set of key variables and recommendations on predicting the probability of an applicant defaulting on their loan.*

## 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:**

The audience consists of the client’s information during the current loan application including demographics and loan details. There are also 6 additional supporting files that contain the applicant’s credit history submitted to Credit Bureau, previous application details, previous POS and cash loans, and installment payments.

**Observation window: N/A**

**Inclusions: N/A**

**Exclusions:** First time clients, no data in credit bureau

**Data Sources:** Kaggle – Home Credit

**Audience Level:** Professional

**Variable Selection:** To be determined.

**Derived Variables:** To be determined.

**Assumptions and data limitations:** To be determined.

## 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 / Mitigation** |
| Highly imbalanced dataset | *High* |  | Only 8% of the dataset is classified as default. High imbalance of the non-default class might lead to a biased model towards it; therefore, producing an inaccurate result. In order to mitigate this, oversampling or undersampling techniques should be applied. |
| Skewed data | *High* |  | Dataset is not normally distributed for some variables which can create an unreliable result. This can be mitigated by performing transformation techniques that will normalize the data. |

## 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 | This project will be formally initiated by sending a request to the advisory team. |  | 5 | July 8, 2023 |
| 2. | Assessment / Triage | Review the dataset and identify data cleaning activities needed to prepare the data for modelling.  Exploratory data analysis will be conducted to check on the data structure, missing values, skewness, outliers, and duplicates. This step will yield insights into the data distribution and target variable characteristics.  Data issues will then be triaged and prioritized based on importance. |  | 6 | *July 13, 2023* |
| 3. | Prioritization |  | *July 13, 2023* |
| 4. | Data Exploration & Analysis   * Issues with duplicates * Issues with Spend data |  | *July 13, 2023* |
| 5. | QA Output | Address data quality issues identified in step 4. |  | *6* | *July 24, 2023* |
| 6. | Data Processing & Feature Engineering | Merge and create aggregated variables from other supporting files to the main table (application\_df) based on SK\_ID\_CURR. Some categories will be merged and create new variables if needed. |  | *July 24, 2023* |
| 7. | Modeling and evaluation | Create predictive models, evaluate performance, and adjust if necessary to improve accuracy. |  | *14* | *July 25, 2023* |
| 8. | Insight generation and presentation |  |  | *5* | *August 8, 2023* |
| 9. | Documentation |  |  | *August 8, 2023* |
| 10. | Advisory Team Presentation |  |  | 1 | *August 14, 2023* |

# References

FICO. 2022. *Using Alternative Data in Credit Risk Modelling*. <https://www.fico.com/blogs/using-alternative-data-credit-risk-modelling>

Home Credit. <https://www.homecredit.net/about-us.aspx/>

Kagan, J. 2023. *Default Risk: Definition, Types, Ways to Measure*. Investopedia. <https://www.investopedia.com/terms/d/defaultrisk.asp#:~:text=Default%20risk%20refers%20to%20the,credit%20reports%20and%20credit%20scores>.

TransUnion. 2023. *More than 9 million Canadians are either credit unserved or underserved; approximately 14% migrate to being credit active every two years.* <https://newsroom.transunion.ca/more-than-9-million-canadians-are-either-credit-unserved-or-underserved---approximately-14-migrate-to-being-credit-active-every-two-years/#_edn2>