Analytics Startup Plan

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

Project	*Analysis of Non-Payment Probability for Canadian Households*
Requestor	*Centennial College*
Date of Request	*July 6, 2022*
Target Quarter for Delivery	*August 10, 2022*
Epic Link(s)	*Not Applicable, we don´t have an agile group*
Business Impact	*Create insights to assess the probability of non-payment*

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.

The specific ask:

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

This project will explore the probability of non-payment for households across Canada. Its findings are expected to serve as a barometer of the Canadian economy and yield valuable insights into the optimization of business practices.

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?

Most institutions, private and public, base their short-term operations on what they perceive to be the current state of the economy, while expectations about the economic cycle largely influence long-term business plans. Ultimately, many current and future business decisions depend on consumers' financial wellbeing, employment conditions, and perceptions about their own future.

This project has been designed to help businesses, regulatory bodies, and/or consulting agencies:

- improve their short-term strategies by defining the characteristics of the households with the highest and lowest probabilities of skipping or delaying a payment
- design sound long-term strategies by providing a baseline of the "financial health" of Canadian households

Technical note: this project may potentially be used as the starting point for additional models. For example, by replicating this model with new data, an analyst could compare the financial strength of households of interest versus a previous period. A clustering analysis could be performed later to look into the relationship between previously defined household groups and a specified target variable.

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 findings of this project have the potential of helping businesses, regulatory bodies, and/or consulting agencies:

- reduce their exposure to client default risk
- reduce their cost of lending/credit
- improve their profitability
- design short and long-term business plans
- gain a better understanding of the financial wellbeing of the Canadian population

The opportunity costs of this project potentially include additional costs of lending, missed revenue, and poor policy construction.

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 objective of this analytics project is to create at least one model that accurately predicts if a household will skip or delay a non-mortgage payment (target variable)
- Due to the nature of the business problem at hand, the interpretability of the predictive model that will be selected is really important. This situation limits, to some extent, the methodologies that may applied during the completion of this project
- One of the main tasks is finding what variables increase or decrease the probability of getting a positive response to the target variable, and how these interactions occur
- The probability of delaying or skipping a payment is expected to be related to demographics, income, and wealth. The direction of these interactions has not been theorized in order to mitigate the potential of introducing bias to the analysis
- Given that the target variable is binary, the models that will be created fall into the category of classification

2.1 Other related questions and Assumptions:

List any assumptions that may affect the analysis

- Other than inheritance, there are no variables that directly capture the effects of phenomena like human capital transmission (academic grade of parents, socio-economic status of parents' household, etc.) or macroeconomic conditions (gross domestic product, inflation, unemployment rates, etc.). However, these phenomena may indirectly play a significant role in explaining the target variable. The potential of serial correlation exists for some of the models that will be created.
- It should be assumed that the dramatic disproportion of wealth and income found across Canadian households may lead them to behave considerably differently. In other words, the reasons that lead the ultra-wealthy households to miss or delay a payment may be completely different to those for the rest of the population. This situation introduces the potential of finding seriously skewed data.

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.

Measures of success include:

- Default rates for banking institutions*
- Account receivables turnover
- Days of sales outstanding
- Bad debt ratio
- Total cost of lending/credit
- Success rates for the launch of new products [long-term business plans]

*In the banking industry, high credit risk has been associated to high emissions. By reducing credit risk, institutions can claim to be taking an environmentally responsible stance.

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 Tree, linear regression, full regression, forward inclusion regression, backward exclusion regression, stepwise regression, and neural networks.
A classification tree will be used as the starting point to determine what are the variables that play the most significant role in explaining the target variable. Secondly, a set of regressions will be run as a variable selection method. A set of neural networks will be run on the full dataset. Lastly, an additional set of neural networks will be run on the variables returned by the regression with the highest accuracy metrics.

Methodology: During the initial phase of this project, a number of data cleaning tasks that include categorical variable class recoding, imputations, exclusions, cap and floors, transformations, and tests of multicollinearity, correlation, and skewness will be completed.
The second phase of this project will mostly include simple modeling tasks. Trees and a full regression will be created to obtain a preliminary outlook of variable importance.
The third phase will include the implementation of additional analytics models like logistic regressions and neural networks. The models created during this final phase are expected to yield the best accuracy metrics. As it has been previously stated, the analytics objective of this project is to generate valuable insights related to the variables that have the greatest impact on the probability that a household will skip or delay a payment. For this reason, the interpretability of the selected model must be taken into consideration.

Output: the output of this project will aid in the creation or improvement of accurate descriptions of households that have the highest probability and lowest probabilities of skipping or delaying a payment. Secondly, if this model is replicated in later years, this project should serve as the starting point to assess the Canadian households' financial strength over time.

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: Canadian households

Observation window: 2018 - 2019

Inclusions: None

Exclusions: Ultra-wealthy Canadian households (as defined by the number of standard deviations away from the population mean).

Data Sources: Statistics Canada: Survey of Financial Security 2019

Audience Level: All Canadian households

Variable Selection: Logistic regressions will be implemented as the main method of variable selection for this project.

Derived Variables: Dummy variables will be systematically created for different dimensions of wealth original captured by dollar-amounts data. For example, if a household has one dollar or more invested in a given financial instrument, a value of 1 shall be assigned to this household on the corresponding dummy variable.

Assumptions and data limitations: Data is complete. Skewness is expected in many variables due to the income/wealth nature of this survey. Additionally, no variables of related to the economic cycle or human capital transmission are included.

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
Working with a population dataset introduces the risk of creating inefficient estimators due to stark contrasts among them.	*Medium*		*It is expected that the basis model of this project may not be as efficient. However, based on our initial findings, a clustering analysis may be constructed to find more reliable results across populations with distinctly similar behaviors.*

Data may not be stationary if this model is replicated in the future.	*Medium*	*Medium*	*The accuracy of this model may not hold in the future if the characteristics of households change significantly and precautions are not taken.*

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	*The project will be requested and approved. The starting date is expected to be August 4.*	*Approval of project and data*	*4*	*July 4*
2.	Assessment	*Initial consultation with advisor. Approval of objectives, early discussions about dataset. Analysis of data dictionary.*	*Assessment of data dictionary*	*2*	*July 6*
3.	Prioritization	*Discussions about early challenges. Any problems with data? What should be the first steps to complete EDA and modeling on time?*	*Definition of tasks and importance*	*2*	*July 8*
4.	Data Exploration & +Analysis +Issues with duplicates Issues with Spend data	*Data quality will be assessed and problems corrected. Formal exploration of dataset.*	*Data quality assessment – data exploration*	*10*	*July 22*
5.	Story Board 1	*Modeling of data – discussion of early model findings. Do early results align with expectations?*	*Early insights*	*2*	*July 26*
6.	QA Output	*Revisit data cleaning/feature engineering section? Revisit modeling objectives.*	*Revisit early objectives*	*3*	*July 29*
7.	Documentation	*Preparation of data governance plan and documentation.*	*Documentation*	*4*	*August 4*
8.	Story Board 2	*Discussion of main challenges and findings with peers.*	*Peer review*	*1*	*August 5*

9.	Internal Presentation	*Executive presentation to main and secondary advisors*	*Presentation of findings*	*3*	*August 8*
10.	Delivery	*Preparation of final deliverables.*	*Final delivery*	*7*	*August 15*