# Analytics Startup Plan

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

|  |  |
| --- | --- |
| **Project** | *Data-Driven Optimization of Loan Approval Process.* |
| **Requestor** | *Miguel Polanco* |
| **Date of Request** | *2024-07-15* |
| **Target Quarter for Delivery** | *August 2024* |
| **Epic Link(s)** | [Federal Trade Commission (FTC) (kaggle.com)](https://www.kaggle.com/datasets/willianoliveiragibin/federal-trade-commission-ftc) |
| **Business Impact** | *Implementation of an Automated System for real-time predictions for the bank loan approval of clients.* |

## 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 opportunity to be presented is to improve loan approval efficiency, reduce the risk of bad loans and offer personalized loan options to clients. The primary problem aims to solve is the inefficiency and risk in the property loan approval process. The bank currently faces challenges in quickly and accurately assessing loan applications, which can lead to approval delays, increased risk of bad loans, and not optimal loan offers for applicants.

Developing a machine learning model to predict loan approval, can provide the bank with the benefit of Improve loan approval efficiency, reduce risk of bad loans and offer personalized loan options.

**The specific ask:**

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

To achieve this opportunity, it is important to consider the relevant variables for the model and to remove the variables that can produce any correlation between them. It is important to take care of the missing values either by imputing or removing them based on their significance and patterns. Categorical features like Gender, Married, and Education are converted into numerical formats. Also, Numerical features such as ApplicantIncome and LoanAmount are scaled to ensure that everything in the model is run with the same standards. After data preprocessing is completed, decision trees will be run as well as an impute, cap and floor and transformation if needed. After that is time to run some regressions and neural networks. After the models are completed, it is important to evaluate each model and perform model comparison using ROC index and ASE. After that the model with the best score should be selected and interpreted as well as the odds ratio and chi square of it.

## 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? |

Economic indicators such as unemployment rates, inflation, and interest rates influence the ability and willingness of individuals to take on loans. During economic downturns, banks may tighten lending criteria, leading to higher rejection rates. Conversely, in a thriving economy, loan approval rates may increase as borrowers' financial stability improves. Demographic factors such as age, gender, marital status, and education level can significantly influence loan approval rates. Research has shown that certain demographics may face higher rejection rates due to systemic biases or financial instability. Understanding these trends helps in tailoring the model to account for such biases and ensure fair predictions. Employment status and income levels are pivotal in assessing loan applicants' repayment capabilities. Self-employed individuals or those with irregular income streams might face higher scrutiny compared to salaried employees.

## 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.* |

For this project the revenue gains for the business are the increase in loan approvals. So, by accurately predicting loan approvals, the bank can improve its approval process, leading to a higher volume of approved loans. This directly translates to increased revenue through interest and fees on additional loans.

For quality improvement the business can create a better customer experience and improve risk management. With the creation of this model, the ability to quickly and accurately process loan applications will benefit the overall customer experience. Faster approvals and personalized loan offers will increase customer satisfaction and loyalty. Also, accurate prediction of loan approval helps in better assessment of risk profiles, reducing the likelihood of defaults. This improves the quality of the bank’s loan portfolio and its overall financial health.

For cost and time savings, this opportunity will reduce the processing time and cost by the automatization of the process.

This approach will benefit the customers by providing faster approvals, better loan terms and by optimizing the customer service experience.

The implications of doing nothing are the loss of a potential revenue opportunity, customer dissatisfaction, higher risks of default and increased operational costs.

## *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. |

What factors most significantly influence loan approval decisions?

Which demographic, employment, loan details, and credit history features are most predictive of loan approval?

How will faster loan approval times affect customer satisfaction and retention?

What are the risks associated with deploying an automated loan approval model, and how can they be mitigated?

How will the predictive model align with the bank’s strategic goals and KPIs?

What cost savings can be achieved through automation of the loan approval process?

How will the implementation of a predictive loan approval model impact the bank’s revenue?

The dataset is assumed to be accurate, representative, and free from significant errors or biases.

The features included in the dataset (demographic, employment, loan details, and credit history) are assumed to have predictive power for loan approval.

Each loan application is assumed to be independent of others, with no hidden dependencies affecting the predictions.

The patterns and relationships in the historical data are assumed to be stable and applicable to future loan applications.

Implementing the predictive model will increase loan approval rates, resulting in higher revenue.

Null Hypothesis (H0): The predictive model will not significantly impact loan approval rates or revenue.

Alternative Hypothesis (H1): The predictive model will significantly increase loan approval rates and revenue.

## 2.1 Other related questions and Assumptions:

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

## 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.* |
|  | To select the successful model, the key performance indicators will be the ROC Index and ASE score as a second key performance indicator. The best model will be the one with higher ROC and lower ASE. |
|  |  |

## 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:** The primary objective is to build a model to accurately classify applicants as approved or rejected based on these features. This involves using most of the features in the dataset to achieve this goal and integrating the model into the bank's loan processing system for real-time predictions. The techniques used to complete this objective would include the use of decision trees, regression models and neural networks.

*logistic regression, linear regression, Chi-square test*

*The initial approach will be to use a decision tree to determine which dealer level variables (size, region, segmentation...) are most significant related to a dealer’s likelihood to churn. We will also use other techniques to verify our findings.*

**Methodology:** The project will first identify the target variable which will be the loan\_status. After performing all the data preprocessing, decisions trees will be built to check which variables are the most important for the objective of the project. After this, it is important to check if imputing is necessary to reduce skewness in some variables. If necessary, also a cap and floor can be used to reduce skewness in the variables. If this still doesn't reduce the skewness for the variables, then a log transformation should be done to the affected variables. Then, with this data it is possible to start the regression models and neural networks. A full regression model, forward regression model, backward regression model and stepwise regression model will be made. Also, a neural network of all the trees, imputes, transformations, cap and floors and regression models will be run. If a neural network is found to be the one with the lowest ASE, then will be time to run 8 more neural networks based on the neural network with the lowest ASE and changing the number of hidden units from 1 to 8. Finally, a model comparison will be done to determine which model is the best based on their ROC index. Interpretations of each aspect of the process and for the best model would be made to develop the recommendations for the bank.

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

*We will start by identifying all dealers that were active in the first quarter of 2018. We will then define the response variable to be a 1 if they are still active, and 0 otherwise. We will build a decision tree based on this sample, and observe which variables are the most important in determining whether these dealers are still active. We can then repeat this analysis using a sample based on the dealers that were active in the second quarter of 2018. The idea is to check if the same variables are being identified as the most important drivers of churn, or if the importance of variables change as we get closer to the present day.*

**Output:** The output will be a set of recommendations that will help the bank to accurately classify applicants as approved or rejected and to integrate the model into the bank's loan processing system for real-time predictions.

*The output will be a set of insights, rules and strategic recommendations that will help us to evaluate dealers based on likelihood to churn and positioning of sales-match.*

## 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:** 381 different loans

**Observation window:** Past applicants considering variables

**Inclusions:** Loan\_Status

**Exclusions:** Loan\_ID

**Data Sources:** [Federal Trade Commission (FTC) (kaggle.com)](https://www.kaggle.com/datasets/willianoliveiragibin/federal-trade-commission-ftc)

**Audience Level:** Each loan\_ID will be an individual observation

**Variable Selection:** Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, Credit\_History, Property\_Area

**Derived Variables:** Loan\_Status

**Assumptions and data limitations:** The dataset is assumed to be accurate, representative, and free from significant errors or biases. The features included in the dataset (demographic, employment, loan details, and credit history) are assumed to have predictive power for loan approval. Each loan application is assumed to be independent of others, with no hidden dependencies affecting the predictions. The patterns and relationships in the historical data are assumed to be stable and applicable to future loan applications.

For data limitations: Significant differences in the distribution between approved and rejected loan applications can bias the model towards the majority class. Insufficient historical data on loan applications may restrict the model's ability to generalize to future applications. If the dataset is not representative of the population or if there are biases in the sampling method, the model may not generalize well.

## 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** |
| *Poor-quality data, missing values, and limited historical records can lead to inaccurate model predictions.* | *Medium* | *Medium* | *Implement rigorous data cleaning, validation, and imputation techniques.* |
|  |  |  |  |
| *The model may unintentionally exhibit bias against certain educations levels, leading to unfair loan approval decisions.* | *Low* | *Low* | *Conduct thorough fairness and bias analysis. Use techniques to detect and mitigate bias, ensuring the model adheres to ethical standards and regulatory requirements.* |
| *Challenges in integrating the predictive model into the bank’s existing loan processing systems can delay implementation and reduce efficiency.* | *Medium* | *High* | *Plan for seamless integration by involving IT and operations teams from the start. Ensure compatibility with current systems.* |
| *Complex machine learning models may lack transparency, making it difficult for stakeholders to understand and trust the model’s decisions.* | *Medium* | *Medium* | *Prioritize the use of interpretable models (e.g., logistic regression, decision trees).* |

## 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 | *Start of the project and approval of it.* | *100%* | *1* | *2024-07-03* |
| 2. | Assessment of topic. | *Discussion of topic selection and questions about Analytics Plan with the advisor.* | 100% | 5 | *2024-07-08* |
| 3. | Analytics Plan for Project | Completion and submission of the Analytics Plan. | 100% | 6 | *2024-07-15* |
| 4. | Data Exploration & Analysis   * Issues with duplicates * Issues with Spend data | Revision of the data, variables and target variables. Relationship between variables. Review of data structure, data quality and variables handling. | 100% | 7 | *2024-07-22* |
| 5. | Modeling | Executing the project models. (Decision Trees, Regressions, Neural Network) | *100%* | *14* | *2024-08-05* |
| 6. | Review of modeling and interpretation | Review of results, interpretation and analysis of the data. | *100%* | *2* | *2024-08-07* |
| 7. | Documentation | Finalize the documentation of the findings for the project. | *100%* | *7* | *2024-08-15* |
| 8. | Presentation | Presentation of the project. | *100%* | *7* | *2024-08-15* |
| 9. | Submission of project | Submission of final project. | *100%* | 1 | *2024-08-16* |