# Analytics Startup Plan

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

|  |  |
| --- | --- |
| **Project** | *Building a Prediction model for banking loan approval* |
| **Requestor** | *Nationalized banks* |
| **Date of Request** | *August 28th 2023* |
| **Target Quarter for Delivery** | *Q3(Third quarter)* |
| **Epic Link(s)** |  |
| **Business Impact** | *The project aims to address the challenge faced by the bank in accurately determining the creditworthiness of loan applicants. By developing a prediction model, the bank aims to improve the loan approval process and minimize the risk of granting loans to undeserving applicants. This will help the bank make more informed decisions, reduce potential losses, and enhance overall profitability. Additionally, the project will enable the bank to streamline its operations and provide better customer service by automating certain aspects of the loan approval process.* |

## 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. |

Business Statement: The business opportunity is to develop a prediction model for a nationalized bank to accurately determine loan approval or rejection for loan applicants.

Opportunity/Problem: Currently, the bank faces challenges in identifying deserving loan applicants out of a large pool of applicants. Although a rigorous verification and validation process is followed, there is no certainty in selecting the right applicants for loan approval. This uncertainty poses risks to the bank's assets and profitability. Therefore, there is a need to leverage machine learning techniques to build a prediction model that can assess the creditworthiness of loan applicants more accurately.

By addressing this problem, the bank aims to streamline its loan approval process, reduce the risk of granting loans to undeserving applicants, and improve overall profitability. The prediction model will help in automating certain aspects of the loan approval process, enabling the bank to make more informed decisions and provide better customer service.

Communicating the Business Opportunity: The nature of the business opportunity, along with the problem statement and desired outcomes, will be communicated to all stakeholders involved in the loan approval process. This will ensure alignment and a shared understanding of the project's objectives. Regular updates and progress reports will be provided to keep stakeholders informed throughout the project's development and implementation stages.

## 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? |

There are certain aspects that are commonly considered during the loan review process. I will explore the documentation, financial projections and narratives that are typically required when applying for a loan. Additionally, I will discuss the challenges faced by lending institutions in 2023.

Key Aspects of Loan Applications:

When applying for types of loans such as home equity credit, business working capital, short term loans, equipment loans or real estate financing there are fundamental lending principles that come into play. These principles include:

1. Credit history: Banks carefully examine the borrowers credit history to assess their creditworthiness and determine their ability to repay the loan.

2. Cash flow history and projections: Banks analyze the historical cash flow of a business as projected future cash flows to evaluate its capacity to repay the loan.

3. Collateral availability: The presence of collateral that can be used to secure the loan is a factor for Banks to consider.

4. Character: Banks also assess the borrowers character by considering factors, like reputation, integrity and business expertise.

By focusing on these areas borrowers can enhance their chances of securing loan terms while lenders can make informed decisions based on a thorough evaluation process.

Challenges Encountered by Nationalized banks / Lending Institutions:

In 2023 lending institutions face obstacles that require navigation. These challenges include:

Ensuring Accurate Loan Decision-Making Amidst Time Constraints:

Nationalized banks are under increasing pressure to match the swiftness offered by private counterparts in loan origination. This urgency is further compounded by the disruption caused by the COVID-19 pandemic, which has necessitated remote processes, impacting traditional in-person assessments. However, maintaining accuracy in loan approvals or rejections within such compressed timelines remains a challenge.

Struggling with Delays and Inefficiencies in Loan Approval:

The complexity of the loan origination process often involves multiple parties, introducing potential bottlenecks and inefficiencies. While online lending platforms provide a digital ecosystem for stakeholder interaction, the challenge persists in ensuring that this streamlining truly expedites the approval process, minimizing delays and enhancing overall operational efficiency.

Balancing Compliance Amidst Regulatory Divergence:

Navigating divergent regulations presents a substantial challenge for nationalized banks. Despite the potential assistance offered by digital tools in organizing compliance efforts, the intricate landscape of rules and regulations can still lead to confusion and potential mistakes. Effective utilization of resources, forums, and surveys becomes paramount to maintain compliance and avoid costly errors.

Mitigating Costs while Ensuring Sound Loan Decisions:

Embracing cost-effective methods for loan origination is a strategic necessity for nationalized banks. The integration of digital solutions can indeed optimize efficiency and reduce operational overheads. However, the challenge lies in striking a balance between cost optimization and ensuring that these streamlined processes do not compromise the accuracy and soundness of loan decisions, ultimately impacting the bank's financial health and reputation.

Overall lending institutions must recognize these challenges. Embrace solutions to stay competitive in an evolving landscape.

While there may be expenses involved in setting up platforms and providing employee training online lending can yield returns, on investment (ROI) by lowering origination costs boosting productivity and harnessing the power of AI technology.

Reference

Livevox. (2023, January 30). *Top Challenges for Lending in 2023*. LiveVox. <https://livevox.com/top-challenges-for-lending/#gref>

Peterson, D. (2018, November). *How Automation Can Improve Your Loan Origination Process*. Www.moodysanalytics.com. <https://www.moodysanalytics.com/articles/2018/maximize-efficiency-how-automation-can-improve-your-loan-origination-process>

## 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.* |

Revenue Gains:

Implementing a prediction model for loan approval/rejection can lead to significant revenue gains for the nationalized bank. By accurately predicting whether a loan should be approved or rejected, the bank can minimize the risk of default and focus on granting loans to deserving applicants. This can result in reduced bad debt expenses and improved overall loan portfolio performance. With a lower default rate, the bank can attract more customers and generate higher interest income from successful loans.

Quality Improvements:

The prediction model can enhance the quality of loan approval decisions by incorporating various factors and historical data. It can identify patterns and correlations that may not be apparent to human underwriters, leading to more accurate predictions. By reducing the likelihood of granting loans to high-risk applicants, the bank can maintain a healthier loan portfolio with lower default rates, ultimately improving the overall quality of the loan book.

Cost and Time Savings:

Implementing a prediction model can save considerable costs and time for the bank. Traditionally, loan approval processes involve manual verification and validation, which can be time-consuming and resource-intensive. By automating the decision-making process using the prediction model, the bank can reduce the need for manual intervention and accelerate loan approval cycles. This can result in cost savings by reducing the number of staff required for loan processing and increasing operational efficiency.

Customer Satisfaction:

Customers will benefit from the implementation of a prediction model for loan approval/rejection. The model can provide more objective and consistent loan decisions, ensuring fairness and transparency in the approval process. Deserving applicants are more likely to receive loan approval, which can enhance customer satisfaction and loyalty. Moreover, faster loan approval cycles can improve the overall customer experience, making the bank a preferred choice for loan applicants.

Implications of Inaction:

If the bank chooses not to develop and implement a prediction model for loan approval/rejection, it may face several implications. First, the bank may continue to approve loans without a reliable method to assess the creditworthiness of applicants, leading to higher default rates and increased bad debt expenses. This can adversely impact the bank's financial performance and profitability. Second, manual loan processing can result in longer approval cycles, causing delays and frustration for customers. Third, without leveraging advanced data analytics and machine learning techniques, the bank may lag behind competitors who have embraced predictive modelling for loan approvals, potentially losing market share.

In conclusion, developing a prediction model for loan approval/rejection can bring significant revenue gains, quality improvements, cost and time savings, and enhanced customer satisfaction for the nationalized bank. Failing to adopt such a model may result in increased risks, operational inefficiencies, and diminished competitiveness in the market. Therefore, prioritizing this project aligns with the bank's goals and key performance indicators.

## *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. |

Key Questions:

1. What factors contribute to the approval or rejection of a loan application?
2. Can we accurately predict whether a loan application will be approved or rejected based on the available variables?
3. Which variables have the most significant impact on loan approval/rejection?
4. Can we identify any patterns or trends in the data that may be helpful in understanding loan approval outcomes?

Assumptions:

1. The provided variables are sufficient to build a predictive model for loan approval/rejection.
2. The dataset contains a representative sample of loan applications from the target population.
3. The historical loan approval/rejection data is reliable and accurate.

Hypotheses:

1. Applicants with higher CIBIL scores are more likely to have their loan applications approved.
2. New customers (NEW\_CUST = 'Y') may have a lower probability of loan approval compared to existing customers.
3. Salaried employees (EMPLOYMENT\_TYPE = 1) may have a higher likelihood of loan approval compared to self-employed individuals (EMPLOYMENT\_TYPE = 0).
4. Age may play a role in loan approval, with younger applicants having a lower probability of approval.
5. Loan applicants with a higher number of dependents (NO\_OF\_DEPENDENTS) may face more scrutiny and have a lower probability of loan approval.
6. Marital status (MARITAL\_STATUS) may influence loan approval, with married applicants having a higher likelihood of approval.
7. Educational qualifications (EDU\_QUA) may positively impact the probability of loan approval.
8. Applicants with a lower income-to-expense ratio (INCOME\_EXP\_GMI) are more likely to have their loans approved.
9. The type of employment (EMPLOYEE\_TYPE) may influence loan approval, with government employees (EMPLOYEE\_TYPE = 2) having a higher likelihood of approval.
10. The loan tenure (TENURE) may affect the probability of loan approval, with longer tenures being associated with higher approval rates.

Deliverables:

1. A predictive model for loan approval/rejection based on the available variables.
2. Evaluation metrics (AUC-ROC, F1-score) to assess the model's performance.
3. Identification of the most influential variables in the loan approval process.
4. Insights and recommendations based on the analysis to improve loan approval outcomes.
5. Optionally, a recommended operating model or blueprint for implementing the predictive model into the bank's loan approval process.

## 2.1 Other related questions and Assumptions:

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

When considering the assumptions that may influence the analysis, for building a loan approval prediction model it is important to pay attention to the following factors:

1. Data Quality: Assumptions are made based on the assumption that the provided data's accurate, complete and reliable. If there are any concerns about data quality steps should be taken to address and rectify any inaccuracies or missing data to ensure the prediction models performance and validity.

2. Relevance of Variables: It is assumed that the variables provided in the dataset are relevant for predicting loan approval or rejection. However it is essential to consider whether these variables encompass all information or if there might be relevant variables missing from the dataset.

3. Linearity: The assumption of linearity is made, assuming that a linear model can effectively represent the relationship between predictor variables and loan approval status. Nonetheless it should be acknowledged that non linear relationships may exist and may not be adequately captured by a model.

4. Independence of Observations: It is assumed that each observation in the dataset is independent of others. However if there are dependencies or correlations between observations this assumption may be violated. Careful consideration should be given to addressing any dependencies or correlations to ensure analysis.

5. Absence of Multicollinearity: The assumption is made that predictor variables are independent of each other without multicollinearity issues. To avoid multicollinearity it is crucial to assess correlations between predictor variables. Multicollinearity can impact model performance. Make it difficult to interpret the importance of variables.

6. Data Generalization: The prediction model is assumed to generalize to unseen data. However it is essential to validate the models performance, on datasets to ensure its accuracy and reliability when applied in real-world scenarios beyond the dataset.

By taking these assumptions into account during the analysis it helps to establish a reliable loan approval prediction model. I can reasonably infer that the patterns and relationships acquired from the given dataset can be reliably extended to loan applications.

## 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 define success measures/metrics for the loan approval prediction model, it's important to consider the objectives of the project and the expectations of key stakeholders. Here are some potential success measures and key performance indicators (KPIs) that will be used to evaluate the model:   1. Accuracy: The overall accuracy of the prediction model in correctly classifying loan approvals and rejections.    * KPI: Accuracy rate (%) 2. Precision and Recall: Assessing the model's performance in identifying loan approvals and rejections accurately.    * KPIs: Precision (%), Recall (%) 3. F1 Score: A combined measure of precision and recall that balances both metrics.    * KPI: F1 Score 4. AUC-ROC: Evaluating the model's ability to discriminate between loan approvals and rejections.    * KPI: Area under the ROC curve (AUC) 5. True Positive Rate (TPR) and False Positive Rate (FPR): Examining the model's performance in correctly identifying loan approvals and minimizing false positives.    * KPIs: TPR (%), FPR (%) 6. Feature Importance: Determining the relative importance of different variables/features in predicting loan approvals or rejections.    * KPI: Importance scores or ranking of features 7. Model Training and Prediction Time: Assessing the efficiency of the model in terms of training time and prediction time for new loan applications.    * KPIs: Training time (seconds), Prediction time (seconds) 8. Generalization: Evaluating how well the model performs on unseen data or a holdout/validation dataset.    * KPI: Validation accuracy (%) |
|  |  |

## 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: For the given project of building a prediction model for banking loan approval, we can employ the following types of analyses:**

1. **Random Forest: This is a suitable analysis technique for binary dependent variables, such as the loan approval status in your case where it's either 1 (approved) or 0 (rejected). Random Forest works by creating an ensemble of decision trees, where each tree is trained on a different subset of the data and makes its own prediction. The final prediction is then determined by combining the predictions of all the individual trees.**
2. **Regression Tree (Decision Tree): Regression trees are useful for exploring non-linear relationships between variables. They can help identify the most significant variables and their interactions in determining loan approval. By constructing a decision tree based on the provided variables, we can observe the importance of different features in predicting loan approval/rejection.**
3. **XGBoost: XGBoost is a robust and versatile machine learning algorithm renowned for its performance in predictive tasks. It excels in capturing intricate relationships within data and delivering accurate predictions. By employing an XGBoost model, we can meticulously train and validate data to make precise predictions regarding loan approval using the provided variables.**

**Methodology: To build the prediction model for loan approval/rejection, we can follow the following methodology:**

1. **Data Collection and Preprocessing: Gather a dataset containing the variables mentioned in the project description (APP\_ID, CIBIL\_SCORE\_VALUE, NEW\_CUST, CUS\_ATGCODE, EMPLOYMENT\_TYPE, Age, SEX, NO\_OF\_DEPENDENTS, MARITAL\_Status, EDU\_QUA, P\_RESTYPE, P\_CATEGORY, EMPLOYEE\_TYPE, MON\_IN\_OCC, INCOME\_EXP\_GMI, ASSET\_LOAN\_RATIO, TENURE, Status). Clean the data by handling missing values, outliers, and categorical variables appropriately.**
2. **Exploratory Data Analysis: Perform exploratory data analysis to understand the distributions, relationships, and correlations between variables. This step will help identify any patterns or insights in the data and guide feature selection for the prediction model.**
3. **Feature Selection: Select relevant features that are likely to have a significant impact on loan approval. This can be done using techniques such as correlation analysis, feature importance from decision trees, or domain knowledge.**
4. **Model Development and Training:**

**a. Random Forest:**

**Create a Random Forest model customized for the loan approval prediction task. Construct an ensemble of decision trees, each trained on a subset of the training data. Adjust hyperparameters like the number of trees, maximum depth of trees, and feature sampling strategy to optimize the model's performance. Train the Random Forest model using the training dataset, enabling it to capture complex relationships between variables and loan approval status. After training, assess the model's predictive prowess using metrics such as accuracy,** AUC-ROC **and F1-score. These metrics will offer valuable insights into the Random Forest model's ability to accurately classify loan applications as approved or rejected based on the given variables.**

**b. Decision Tree:**

**Design a decision tree model tailored to the loan approval prediction task. Develop the tree's structure by selecting appropriate features and splitting criteria that best capture the patterns within the data. Train the decision tree model using the training dataset, allowing it to recursively learn and create decision rules that lead to accurate predictions. Adjust parameters like the maximum depth of the tree to control its complexity and prevent overfitting. Once trained, evaluate the decision tree's performance using metrics such as accuracy,** AUC-ROC **and F1-score. These metrics will provide insights into how effectively the decision tree model predicts loan approval outcomes based on the provided variables.**

**c. XGBoost:**

**Design an XGBoost model specifically tailored for the loan approval prediction task. Craft the architecture of the XGBoost ensemble by configuring the number of boosting rounds, maximum tree depth, and other hyperparameters. Utilize the training data to train the XGBoost model, allowing it to learn intricate patterns and relationships within the dataset. Fine-tune the hyperparameters iteratively to achieve the best possible predictive performance. Once trained, evaluate the XGBoost model's effectiveness using various metrics such as accuracy,** AUC-ROC **and F1-score. These metrics will provide a comprehensive understanding of how well the model performs in predicting loan approval outcomes based on the provided variables**

1. **Model Evaluation and Comparison: Compare the performance of the random forest model, decision tree, and xgboost. Analyze their strengths, weaknesses, and interpretability.**

**Output: The output of the analysis will include:**

1. **Prediction Model: A trained prediction model (logistic regression, regression tree, and neural network) that can predict loan approval or rejection based on the given variables.**
2. **Model Performance Metrics: Evaluation metrics (accuracy, AUC-ROC F1 score, etc.) for each model to assess their effectiveness in predicting loan approval.**
3. **Feature Importance: Identification of the most important variables/features that contribute significantly to loan approval/rejection.**
4. **Insights and Recommendations: Interpretation of the models' results, providing insights into the factors influencing loan approval and recommendations for improving the loan approval process.**

## 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. |

**Population: The population for this project would be the applicants of the nationalized bank who have applied for a loan.**

**Observation Window: NA**

**Inclusions: Include loan applications from all types of customers, both new and existing. Capture data from all available sources such as operational systems, data warehouses, CRM systems, and email marketing systems.**

**Exclusions: SEX and CIBIL\_SCORE\_VALUE shall be excluded**

**Data Sources:** [Banking loan approval info for a Nationalized Bank](https://www.kaggle.com/datasets/isotopek/banking-loan-approval-info-for-a-nationalized-bank?select=loan_data.csv)

**Audience Level: The audience level refers to the technical expertise of the individuals who will be using the prediction model. It is important to consider the level of knowledge and understanding of the audience when designing the model and presenting the results. In this case, assuming the audience consists of data scientists, analysts, and banking professionals with a good understanding of machine learning concepts.**

**Variable Selection: Based on the provided variables, the following variables should be included in the analysis and model development:**

* **NEW\_CUST**
* **CUS\_ATGCODE**
* **EMPLOYMENT\_TYPE**
* **Age**
* **NO\_OF\_DEPENDENTS**
* **MARITAL\_STATUS**
* **EDU\_QUA**
* **P\_RESTYPE**
* **P\_CATEGORY**
* **EMPLOYEE\_TYPE**
* **MON\_IN\_OCC**
* **INCOME\_EXP\_GMI**
* **ASSET\_LOAN\_RATIO**
* **TENURE**

**Derived Variables: NA**

**Assumptions and Data Limitations:** **Assumptions have been listed in Section 2.1. Data**

**Limitations have been listed in Section 4.0**

## 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** |
| *Data Quality: Risk of inadequate or poor-quality data that can affect the accuracy and reliability of the prediction model.*  Feature Selection: Risk of selecting irrelevant or redundant features that may not contribute significantly to the loan approval prediction  Bias and Fairness: Risk of the prediction model exhibiting bias or unfairness towards certain demographic groups, leading to discriminatory loan approval decisions.  Model Deployment and Integration: Risk of facing challenges during the deployment and integration of the prediction model into the existing banking system or infrastructure. | *Medium*  *Medium*  Medium  *Medium* | *Low*  *Medium*  *Low*  *High* | *High Impact: Poor-quality data can lead to inaccurate loan approval predictions, potentially resulting in approving loans for undeserving applicants or rejecting loans for deserving applicants. This can affect the bank's profitability and customer satisfaction.*  Moderate Impact: Including irrelevant or redundant features in the model can introduce noise and decrease the model's accuracy. It may lead to suboptimal loan approval decisions and affect the bank's risk assessment process.  High Impact: If the prediction model exhibits bias or unfairness, it can lead to discriminatory loan approval decisions based on factors such as gender, marital status, or ethnicity. This can result in reputational damage, legal consequences, and loss of customer trust.  High Impact: Difficulties in deploying and integrating the prediction model into the existing banking system can cause operational disruptions and delays in implementing the loan approval process. It may result in inefficiencies, increased costs, and potential loss of business opportunities. |

## 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 | Discussion with stakeholder and advisory team about business and project requirement findings. | * Product vision * Product backlog * Business case document |  | *Jul 4th 2023* |
| 2. | Assessment / Triage | Gathering information to get insight of urgency with regard to project sprint deliverables. | Expected deadlines for key deliverable | 1 day | *Jul 6th 2023* |
| 3. | Prioritization | Objectives and requirement of the project will be prioritize based on importance. | Timelines for each characteristic and issues | 5 days | *Jul 6th 2023* |
| 4. | Data Exploration & Analysis   * Issues with duplicates * Issues with Spend data | Analysis will be done with regard to data quality and feature engineering to get the insight of data distribution across the model and target the variable characteristics. | * Methodologies to be used * Quality of data | 1 day | *Jul 11th 2023* |
| 5. | Story Board 1 | Getting insight of employees and employer opinions and extract particular key areas to focus on. | User stories | *8 days* | *Jul 11th 2023* |
| 6. | QA Output | Followed by step 4, issues regarding data will be addressed and sequentially delivered to the stakeholders. | * Limited to data quality issues as analyzed in step 4 * Quality data | *14 days* | *Jul 18th 2023* |
| 7. | Internal team Presentation | Slides will be presented to demonstrate either the project or key deliverables are meeting the expectations. | * Project Vision * Project Backlog * Definition of done * Product Increment * Burndown chart | *4 days* | *Aug 1st 2023* |
| 8. | Go/No Go | As per the reviews by the team during presentation, lists of go and no go will be carried and followed accordingly. | Lists of suggestion or newer issues and priorities | *9 days* | *Aug 5th 2023* |
| 9. | Story Board 2 | Explanation or reviews of system from the end user. |  |  | *Aug 14th 2023* |
| 10. | Pilot | Final delivery will be made after implementing project requirements and addressing new issues raised during previous deliveries. | * Analytics Plan   Pilot system |  | *Aug 14th 2023* |
| 11. | Delivery & sign-off |  |  |  | *Aug 14th 2023* |