BIN381: Milestone 1

Recommender system

5th October 2024

Student Name	Surname	Number
Nosipho Precious	Donkrag	577354
Tshepang	Mokgosi	577685
Pitsi	Pitsi	577216
Nontsikelelo Sharon	Buhlungu	577878

Contents

Business Problem

The main business problem is to develop an intelligent recommender system that can accurately classify customers as eligible or ineligible for a specific service offering. The service provider wants to ensure that the eligibility decision is based on a combination of demographic and financial data. The system will streamline decision-making, support strategic planning, and improve customer satisfaction by making efficient and accurate predictions.

Objective

The goal of this project is to build a classification model based on customer demographic and financial data that will:

- Predict whether customers are eligible for the service.
- Provide insights through data visualization to allow for better strategic decisions.
- Ensure that both technical and non-technical stakeholders can access and understand the system results through easy-to-use dashboards and reports.

Key Objectives:

- Develop a machine learning model to classify customer eligibility based on demographic data.
- Improve decision-making by presenting eligibility outcomes in a user-friendly dashboard.
- Ensure data quality by identifying and handling missing or noisy data.
- Support scalability, reliability, and maintainability of the system for long-term use.

Scope:

The scope of the project includes:

- Data mining and model development: Applying machine learning techniques on customer data to identify eligibility.
- Data preparation and quality assessment: Cleaning, transforming, and ensuring data consistency.
- Visualization and reporting: Creating dashboards and visual aids for analysis, reports for management, and real-time predictions for decision-making.

Exclusions:

- External system integrations (e.g., ERP or CRM systems) beyond what is required to make eligibility predictions.
- Non-technical features that are unrelated to the core data processing and classification task (e.g., customer interaction interfaces).

Stakeholder analysis:

Stakeholder matrix:

Stakeholder	Role	Interest	Influence	Power	Engagement Strategy
Project Sponsor (CEO)	Provides funding and project direction	Wants the project to drive revenue by targeting qualified customers and improving service offerings	High	High	Regular status updates, high-level reports, and financial ROI reviews
Marketing Department	Uses model outcomes to design campaigns	Interested in identifying qualified customers to create targeted marketing campaigns	Medium	High	Provide detailed insights on customer segments and patterns
Data Science Team	Develops and tests the data mining model	Interested in building an accurate model that can predict customer qualification with high precision	High	Medium	Regular feedback loops on data quality and model performance
IT Department	Provides infrastructure and data support	Ensures the project has sufficient technical resources (e.g., data pipelines,	Medium	Medium	Ensure technical requirements are met and data infrastructure is robust

		storage, computational power)			
Sales Department	Uses the results to inform sales strategies	Interested in knowing which customers qualify for specific services to increase conversion	Medium	Medium	Regular updates on customer qualification data for targeting
Legal and Compliance Team	Ensures data privacy and legal compliance	Interested in ensuring the project complies with regulations (e.g., GDPR, POPIA) when handling customer data	High	High	Regular reviews of data privacy and compliance strategies
Customer Support Team	Supports qualified customers post-service offering	Interested in understanding who qualifies for the service so they can provide proactive support	Low	Low	Inform on customer qualification insights for better support
End Customers	Receive or qualify for the service	Interested in receiving relevant and tailored services, while ensuring their data is used responsibly	Medium	Low	Transparent communication on how data is used and the benefits they receive
External Vendors (e.g., data providers)	Provide external demographic or behavioural data	Interested in providing accurate and timely data to support the model's predictions	Low	Low	Ensure data is provided in a timely and compliant manner

User Story:

User Story - Marketing Manager for Credit Qualification

As a marketing manager, I want to analyse customer data to determine which customers qualify for credit based on their income, occupation, and demographics so that I can target marketing campaigns effectively and increase the conversion rate for our credit services.

1. Credit Qualification Insights:

   - Identify qualified customers by analysing factors such as annual salary, gross income, and job stability (e.g., years in occupation, household size, and years of residence).
   - Segment customers by education level, occupation, and marital status to determine how these factors correlate with creditworthiness.
   - Assess risk by examining geographical data (e.g., city, state) to identify regions with lower or higher credit default risks.

2. Income and Occupation Analysis:

   - Compare customers' annual salaries across different occupations to identify those with stable, high-paying jobs that qualify for credit.
   - Analyse salary trends and household size to determine how financial stability impacts credit qualification.
   - Visualize how education levels influence income and credit qualification potential.

3. Geographical and Demographic Insights:

   - Map the distribution of customers across postal codes, cities, and states to identify regions with the highest number of qualified credit candidates.
   - Break down customers by age group, education level, and marital status to determine which demographics are most likely to qualify for credit offerings.

4. Risk and Qualification Filters:

   - Create filters based on income range, occupation, household size, and years of residence to quickly identify customers who meet credit qualification criteria.
   - Visualize customer profiles with high credit potential and group them by salary range and occupation.

5. Expected Outcome:
   - Enable the marketing manager to target high-value customers for credit offerings.

- Provide a detailed, filtered list of customers who qualify for credit based on data-driven analysis, helping increase the efficiency of marketing campaigns.
- Reduce risk by identifying customers with stronger financial backgrounds while excluding those in high-risk areas or with unstable income.

Requirement analysis:

Functional Requirements:

Functional requirements are core features and operations that the intelligent recommender must perform; these are based on the stakeholders' expectations of how the system must perform. Below are the functional requirements for the intelligent recommender system:

- **FR:** Functional Requirement

FR	Title	Description
FR1	Input Data	The system must be able to read and process customer data stored in csv or excel files such as the CustData2.csv.
FR2	Eligibility Prediction	The system must accept individual customer data and have the model identify if a customer is eligible for the service they are applying for or not based on the data it is provided (non-salary factors).
FR	Data processing and cleaning	The system must improve the quality of data before processing by removing corrupt data, handling missing data and standardizing the customer data.
FR3	Features Engineering	The system must generate features suitable for model training and enhancement based on the input data.
FR4	Data Visualization	The system must have an interactive dashboard that allows the credit risk team to drill down, horizontally and vertically into the data of interest. This dashboard must have visual representations of data such as bar graphs, histograms and pie charts (these visual aids must be implemented cognitively).
FR5	ML model	The system must incorporate the use of a machine learning model to classify customers as eligible or not based on the customer data it is provided.
FR7	Deployment	The system must provide an interface suitable for all stakeholders; simple enough for higher management to understand while allowing for complex analysis required by the risk management team such as drilling down to their desired data.
FR8	Reporting	The system must generate reports with visuals for higher management and technical reports such as model performance showing features of importance that affect the eligibility of a customer.

Non-Functional Requirements:

These are the quality features of the system that will focus on the "unseen" attributes of the system. Below are the lists of the non-functional requirements:

Terminology:

- **NFR**: Non-Functional Requirement.
- **Precision**: Measures how many of the customers predicated to be eligible are truly eligible (used to minimise false positives).
- **Recall**: Measures how many of the truly eligible customers did the model predict as eligible (minimise false negatives).
- **F1-score:** a ratio that balances precision and recall (harmonic mean).

NFR	Title	Description
NFR1	Model Performance	The model should aim for an F1-score of at least 75% when predicting and an accuracy score of at least 80%. The model must not take more than 10 seconds when predicting.
NF2	Scalability	The system must be able to scale to accommodate growing customers.
NF3	User interface	The user interface must be simple enough for non-technical users and it must also allow for some complexity for the risk assessment team.
NF4	Maintainability	The system should be easy to update (about twice a year) with new customer data (differential updates) and backed up twice a year (every 6 months).
NF5	Reliability	The system should be consistently available and have an uptime of 99.8%. The system must have mechanisms for handling errors during prediction requests should they arise.

Requirement traceability Matrix:

This is a tool utilized to further analyse the requirements and understand the relationships that exist between the requirements and the success of the project.

ID	Requirement	Type	Priority	Comment
FR1	Input Data	Functional	**High**	This is a core requirement as the system cannot operate if it cannot load the customer data.
FR2	Eligibility Prediction	Functional	**Critical**	This is central function of the system; thus it is critical to the success of the project.
FR3	Data processing and cleaning	Functional	**High**	This will ensure the quality of data is upheld for the ml model.
FR4	Features Engineering	Functional	**Low**	After data analysis the data may not require additional features; this is all dependant on the nature and structure of the data.
FR5	Data Visualization	Functional	**Critical**	The system is supposed to assist in strategic decision making; data visualisation is critical as it is the tool that will be use to communicate among all levels of stakeholders.
FR6	ML model	Functional	**High**	The ml model is the back bone of the recommender system.
FR7	Deployment	Functional	**Critical**	This requirement allows for the access of data by all levels of management.
FR8	Reporting	Functional	**Medium**	The production of actual pdf reports is not critical to the project; because simple and complex analysis will be readily available as the system is deployed through the dashboard.
NFR1	Model Performance	Non-Functional	**Critical**	The model must predict with as high accuracy and precision as possible as to not affect the overall business processes.
NFR2	Scalability	Non-Functional	**High**	This requirement will accommodate the increase in customers should in occur.
NFR3	User interface	Non-Functional	**Medium**	Ensures both technical and non-technical users can use the system.
NF4	Maintainability	Non-Functional	**Medium**	Regular updates will ensure that the system remains relevant to the business' needs and remains reliable.

NFR5	Reliability	Non-Functional	**Critical**	If the system fails it will affect the day-to-day operations of the business, hence reliability is critical.

The analysis highlights the priority levels of each requirement; the critical requirements will affect the success of the project if not met. These requirements include data processing, eligibility prediction, data visualization and deployment all of which are crucial to the success of the system. The non-functional requirements will ensure the longevity and adaptability of the system within the business process.

Success Criteria:

1. **Data Quality & Integrity:**
   - Data Completeness: 95% of the required customer data is complete and available for analysis.
   - Data Accuracy: The accuracy of the input data (e.g., demographic, purchase history) is verified to be at least 98%, ensuring no faulty records affect the outcome.
   - Data Cleaning Process: Ensure that all duplicates, inconsistencies, and outliers are removed or corrected before running the model.
2. **Model Performance:**
   - Prediction Accuracy: The data mining model achieves at least 90% accuracy in predicting which customers qualify for the service.
   - Precision and Recall: The model has a precision rate (correct positive predictions over all positive predictions) of 85% and a recall rate (true positives over all actual positives) of 80%.
   - Model Interpretability: Ensure that the model results can be interpreted by business stakeholders, and key patterns are explainable.
3. **Business Alignment:**
   - Service Qualification Rate: The project correctly identifies at least 80% of eligible customers who qualify based on pre-defined business rules (e.g., income level, purchase history, location).
   - Reduction in False Positives/Negatives: Limit false positives (customers flagged as qualified but aren't) to below 5%, and false negatives (customers who qualify but are missed) to below 10%.
4. **Efficiency & Timeliness:**
   - Time to Complete Analysis: The full data mining and customer qualification process is completed within the agreed project timeline (e.g., 4 weeks).
   - Automated Process Implementation: The process for identifying qualifying customers is automated and can be updated regularly, reducing manual intervention by 90%.

5. Business Impact:

- Customer Outreach Rate: Post-qualification, 70% of the qualified customers are contacted within two weeks of analysis completion.
- Conversion Rate: At least 20% of the identified qualified customers subscribe to or adopt the service after outreach.
- ROI (Return on Investment): The return on the project investment exceeds the target ROI (e.g., a 3x increase) within 6 months of implementation.

6. Stakeholder Satisfaction:

- Client Satisfaction: Ensure at least 90% satisfaction among internal or external clients (e.g., marketing, sales) with the final list of qualifying customers.
- Error Reduction: The project results in a 50% reduction in errors when compared to previous, manual customer qualification efforts.

Data mining Goals and Success criteria:

Aim: data mining is "Knowledge discovery" thus the aim of this project is to discover any patterns and relationships, through analysis, that may exist in the customer data that will hypothetically and hopefully lead to knowledge discovery. This section will discuss the mining goals for the project and how the success for each goal will be measured.

1. Goals:

Business Understanding Goals

The goal is to understand the business process with the aim of identifying why the recommender model is required. This is to ensure that the system created meets the requirements of the stakeholders.	Steps involved: Stakeholder analysis, requirements analysis and documentation. Continuous communication with stakeholders.

Data Preparation Goals

Data Exploration Data exploration with the aim of understanding the customer data we will be working with.	Steps involved: View the columns, datatypes and summary of data. View the percentage of missing values, corrupt sets, and noisy data. Identify categorical and numeric attributes. Identify any duplicated and outliers that may be present in the data.

Data Cleaning and processing The goal will be to ensure that data is accurate, complete, and consistent.	Handle the identified missing values, noisy data and outliers. Consolidate customer data into a format suitable for data mining. Any data normalisation or scaling will happen here.
Attribute analysis The goal here is to apply statistical analysis (correlation analysis) and visual univariate and multi-variant analysis in order to identify relationships or patterns that may be present.	Univariate analysis of columns in the dataset to view the distribution (histograms, box plots) and counts (bar graphs) of attributes in the customer data. Multivariate analysis between columns (correlation matrix) to identify any relationships present between attributes.

Data Transformation Goal

The goal of data transformation is to transform the data into a format suitable for building the required recommender model. The goal here is to transform the raw data into an optimised and structured format that will enhance the model's performance.	○ Transform categorical features into numeric using encoding methods. ○ Apply normalisation or standardization to scale numeric features appropriately. ○ Drop or remove any unimportant features from the dataset that will not contribute to the training of the mode. ○ Feature engineer any additional features that will help train the model. ○ Aggregate (group) related attributes together.

Model Goals

Model selection The goal is to identify the most suitable model for predicting the customer eligibility. This is a classification problem, thus models suitable for classification will be employed such as the random forest model.	The exploration of the following models will occur: ○ Random Forests, and ○ Logistic Regression. The model that performs best (F1-score) will be chosen.
Deployment The goal here is to successfully deploy and integrate the recommender model into the business process.	Integration of the model within the entire system, providing a seamless transition from data analysis and visualisation to customer prediction.

2. Success criteria:

Goal: Data Exploration

Success Criteria

- All attributes are identified as categorical or numerical.
- Percentage of missing values for each attribute is calculated.
- Duplicated records and outliers are correctly identified.
- A descriptive summary of the data has been generated.

Goal: Data Cleaning and processing

Success Criteria

- Missing values, noise and outliers are handled to ensure data consistency.
- Data is consolidated to the required format.
- There are no unresolved quality issues that would compromise the model training.
- Data is consistent and complete.

Goal: Attribute analysis

Success Criteria

- Relationships and patterns between attributes have been identified.
- The distribution of numeric data has been visualised.
- Insights that will help engineer new features have been gained.

Goal: Data Transformation

Success Criteria

- The data has been successfully transformed and made suitable for the machine learning model (there are no non-numeric attributes present). All categorical data has been transformed to numeric using the suitable encoding methods.

Goal: Recommender Model Goals and Deployment

Success Criteria

- Multiple models have been trained and tested and the best one producing the highest accuracy and F1-score has been selected.
- The model has been successfully integrated into the entire system and provides real time predictions.

    - The system successfully supports strategic decision making and is suitable for stakeholders of different levels.

Inventory and Resources:

Summary

In Milestone 1, the Inventory of Resources will list the tools, datasets, and expertise needed for the project. The Data quality Assessment will identify and resolve data issues like missing values, duplicates, outliers. The CRIPS-DM framework will guide the project, helping you move through the stages of business understanding, data preparation, modeling and evaluation.

Inventory of Resources

This section outlines the tools, datasets, and human resources required to complete the project. We need to list all available assets and identify potential gaps in the resources.

Tools and Software

Data Analysis Tools: R, Python for performing data analysis. R markdown could be used for documenting and sharing results.

Data Visualization Tools: Power BI, Matplotlib/Seaborn in Python for creating dashboards and visualization. Power BI will be useful as it generates interactive insights.

CRISP-DM Framework: A well-defined process for the data mining lifecycle, guiding the project at each milestone.

HUMAN RESOURCE

Data Analysts: Group members responsible for cleaning and analysing the dataset.
Business Experts: Understanding the business context and defining success criteria for the model.
Data Visualization specialists: Creating dashboards and visualization for data insights.

Risk Assumptions and Constraints

- Data Quality Issues: Inaccurate or incomplete data may lead to poor model performance.
    - Mitigation: Implement thorough data cleaning, processing, and quality assessment steps.

- Model Bias: The model might inherit biases from historical data (e.g., demographic biases).
    - Mitigation: Monitor for bias by regularly reviewing model outputs and conducting fairness assessments.
- System Scalability: As the customer base grows, the system may face challenges in managing larger datasets.
    - Mitigation: Ensure the system is built with scalability in mind, with flexible infrastructure.
- Stakeholder Engagement: If stakeholders' expectations are not aligned, there may be conflicts between business objectives and system functionality.
    - Mitigation: Engage in continuous communication with stakeholders to gather clear requirements**.**

Assumptions:

- It is assumed that the provided dataset contains sufficient information to train a robust classification model.
- Stakeholders will be available throughout the project for clarifying business objectives and model requirements.
- All stakeholders will have a minimum level of technical understanding required to interact with dashboards and reports.
- Any data privacy or regulatory constraints related to using personal demographic data will be addressed by the organization.

Constraints:

- Time and Resources: The project must be completed within a limited timeframe and with a defined budget.
- Data Constraints: Data availability is restricted to the provided customer dataset, and there may be limitations in the quality or scope of data.
- Technical Constraints: The system must be designed to integrate seamlessly with the company’s existing infrastructure and tools, such as Power BI for visualizations.

1. Time Constraints:

- **Clarification**: Is there a firm deadline by which the system must be operational? Are there specific milestones that need to be completed by certain dates? How will the timeline impact the depth and breadth of the system's features (e.g., complex vs. simple models, dashboards)?

2. Data Constraints:

- **Clarification**:
    - Are you limited to the **provided customer dataset**, or can additional data sources be incorporated?
    - Are there potential issues with the quality of the data (missing, outdated, incomplete data)?
    - Is the data updated in real-time, or is it static? If static, how often is it refreshed, and can this impact model accuracy over time?
    - Do you have **access to all necessary data features** to make an accurate eligibility prediction, or are some key attributes missing or unavailable?

3. Technical Constraints:

- **Clarification**:
    - Does the project have to use specific technologies (e.g., Power BI for visualization or specific databases)?
    - Are there **hardware or software limitations** within the organization (e.g., limited computational power, access to cloud services for scalability)?
    - How does the project need to **integrate with existing systems**, and are there any limitations in terms of compatibility with other tools or platforms used by the organization?

4. Budgetary Constraints:

- **Clarification**:
    - Is there a **budget limit** for the project?
    - Will there be financial constraints that limit the ability to purchase additional datasets, licenses for software, or advanced computational resources?
    - Are there **ongoing costs** (e.g., for cloud infrastructure, maintenance) that need to be factored in?

5. Regulatory and Compliance Constraints:

- **Clarification**:
    - Are there **legal or regulatory requirements** related to using customer data (e.g., GDPR, data protection laws)?

- How will these regulations influence the collection, storage, and processing of personal data?
- Are there constraints on the **types of data** that can be used for model training, such as sensitive demographic information?

6. Human Resources Constraints:

- **Clarification**:
    - Are there constraints related to the **availability of skilled personnel** to work on the project?
    - Will all necessary roles (e.g., data scientists, developers, and project managers) be available throughout the project's lifecycle?
    - Is there **adequate training** for non-technical users to interact with the final system?

CRISP-DM Approach (**Cross industry Standard Process Data Mining)**

CRISP-DM is the methodology guiding this project.

Phase 1 Business Understanding: The first step is understanding the business objectives and translating them into a data mining problem. The business goal is to develop a classification model that predicts which individuals qualify for a specific service offering.

We need to: Define clear business objectives.
Determine success criteria.
Engage stakeholders to understand their expectations and constraints.

Phase 2 Data Understanding:
The focus here is on exploring the provided dataset.
This include; Collecting initial data.
Understanding each variable.
Identifying data quality issues such as missing values, outliers, and duplicates. Creating visualizations to gain insights into relationships between variables.

Phase 3 Data Preparation:
To prepare it for modelling; The following will be performed:

- Cleaning the data.
- Feature selection.
- Feature transformation.

Phase 4 Modelling:

In the phase, We select appropriate algorithms build your classification model. The CRISP-DM process emphasizes that this phase is iterative—modelling choices may influence the need to revisit data preparation or understanding.

Phase 5 Evaluation:

Evaluate its performance. This involves assessing how well the model meets the business objectives and success criteria.
Use performance metrics like accuracy, precision, recall, or F1 score to judge the model.

Phase 6 Deployment:

The final phase, the model is deployed into a real-world environment.

Data Understanding: