BIN381 MILESTONE 7

Christian Anyimadu, Comfort Hemese, Delight Chipiro, Omphile Tladi

11/6/24

Business Intelligence

Contents

Executive Summary

The Customer Eligibility System project is a strategic initiative to transform LangaSat's approach to determining customer eligibility for its satellite internet services. Previously, LangaSat relied solely on annual salary as the primary eligibility criterion, which often provided a limited view of customer suitability. Recognizing the need for a more comprehensive evaluation, this project leverages machine learning, specifically a Random Forest classifier, to build an intelligent recommender system. This system assesses eligibility based on a broader set of financial and demographic factors, including credit risk indicators, years of residence, and household size, thereby offering a more accurate, consistent, and nuanced assessment.

With an impressive accuracy rate of 99.99%, the system provides fast, dependable, and consistent assessments that significantly improve operational efficiency and decision-making accuracy. Integrating seamlessly with LangaSat's existing CRM system through Shiny Server, the Customer Eligibility System enables real-time assessments directly within business workflows. This integration reduces the time and labor required for manual evaluations and generates valuable insights into customer financial behavior and stability profiles, supporting LangaSat's broader strategic objectives.

The project was developed following the CRISP-DM (Cross-Industry Process for Data Mining) methodology, covering all essential stages from business and data understanding to modeling and deployment. Key findings reveal that by incorporating additional variables beyond annual salary, the model's predictive power is enhanced, making it a robust decision-support tool for customer eligibility determination. Additionally, the project places a strong emphasis on ethical data handling practices to ensure that sensitive customer information is processed responsibly, with measures for privacy, fairness, and transparency embedded into the system.

The successful implementation of the Customer Eligibility System marks a significant advancement for LangaSat. It enables the company to make informed, data-driven eligibility decisions, enhances customer engagement, and strengthens operational workflows. Furthermore, the project lays a foundation for future machine-learning applications and continuous performance monitoring, positioning LangaSat for continued innovation and adaptation to evolving customer needs and market conditions.

Introduction

The Customer Eligibility System project was launched to help LangaSat, a satellite internet provider, improve how it assesses customer eligibility for contract services. Previously, eligibility was determined solely by annual salary, but LangaSat recognized the need for a more comprehensive approach. This project aimed to develop an intelligent recommender model using a Random Forest classifier to evaluate eligibility based on multiple factors, including credit risk indicators, years of residence, and household size. Following the CRISP-DM methodology, the project involved data exploration, model building, performance evaluation, and deployment via a Shiny app integrated with LangaSat's CRM, enabling real-time assessments.

The project's goals were to enhance accuracy in eligibility decisions, streamline operational efficiency, and uphold ethical data practices. By automating and expanding the eligibility assessment criteria, LangaSat gains deeper insights into customer profiles, reduces manual processing time, and strengthens its data-driven approach to service offerings. This report covers the project's objectives, methodology, findings, and recommendations for future development.

Project Background

LangaSat, a company offering satellite internet services on contract terms, initially assessed customer eligibility based solely on annual salary. This limited approach did not account for other factors that could indicate a customer's financial stability and suitability for long-term contracts. Recognizing the potential benefits of a more nuanced eligibility assessment, LangaSat embarked on a project to build an intelligent, data-driven system that evaluates customer eligibility based on a broader range of variables. This approach aims to enhance accuracy, improve customer targeting, and streamline service delivery.

Problem Statement

The current eligibility determination process at LangaSat, which relies only on annual salary, lacks depth and may lead to inaccurate decisions regarding customer eligibility. This narrow approach not only increases the risk of approving unqualified customers but also overlooks potentially eligible customers with strong financial stability. LangaSat

requires a solution that leverages additional variables to provide a more accurate and efficient eligibility assessment, reducing manual processing time and improving consistency in decision-making.

Project Scope

The scope of this project includes the development of a machine-learning-based Customer Eligibility System that integrates multiple financial and demographic variables to determine customer eligibility for satellite internet services. The project will:

1. Collect and preprocess relevant customer data.
2. Develop a predictive model, specifically a Random Forest classifier, to evaluate eligibility.
3. Integrate the model into LangaSat's CRM system through Shiny Server for seamless deployment.
4. Implement monitoring and maintenance protocols to ensure model accuracy and reliability over time.
5. Ensure ethical handling of sensitive customer data throughout the project lifecycle.

Project Objectives

1. **Develop a Recommender Model**: Build an intelligent recommender model using a Random Forest classifier to evaluate customer eligibility based on numerous factors beyond annual salary.
2. **Increase Accuracy and Efficiency**: Enhance the accuracy of eligibility assessments, minimize human error, and reduce the processing time associated with manual evaluations.
3. **Integrate with Existing Systems**: Seamlessly deploy the model into LangaSat's CRM system for real-time eligibility assessments within existing business workflows.
4. **Support Ethical Data Practices**: Implement ethical guidelines for data privacy and fairness, ensuring responsible handling of customer information.
5. **Provide Actionable Insights**: Generate insights into customer financial behavior and stability, supporting LangaSat's strategic decision-making and customer engagement efforts.

Business Understanding

LangaSat aims to enhance its customer eligibility assessment process for its contract-based satellite internet services. The traditional method which relies solely on evaluating a customer's annual salary, has proven insufficient in capturing the full picture of a customer's financial stability and creditworthiness. The primary business goals are:

1. **Improve Accuracy of Eligibility Assessments**: By incorporating additional financial and demographic factors such as credit risk indicators, years of residence, household size, and spending patterns, LangaSat seeks to create a more comprehensive and accurate eligibility assessment model.
2. **Increase Operational Efficiency**: Automating the eligibility assessment process will reduce the time and resources spent on manual evaluations, allowing staff to focus on customer service and other value-added activities.
3. **Enhance Customer Satisfaction**: Providing quicker and fairer eligibility decisions will improve the customer experience, leading to higher satisfaction rates and potentially increasing customer retention and acquisition.
4. **Mitigate Financial Risk**: By identifying high-risk applicants more effectively, LangaSat can reduce the likelihood of defaults or contract terminations, safeguarding the company's revenue streams.
5. **Leverage Data for Strategic Insights**: Utilizing data mining techniques will enable LangaSat to uncover patterns and trends in customer behavior, informing marketing strategies, product development, and risk management.

Success Criteria from a Business Perspective

- **Increased Approval Accuracy**: Achieve at least a 95% accuracy rate in eligibility assessments, reducing false positives (approving high-risk customers) and false negatives (rejecting low-risk customers).
- **Reduced Processing Time**: Decrease the average time taken to assess customer eligibility from manual processing times to automated assessments.
- **Cost Savings**: Lower operational costs associated with manual eligibility assessments through automation.
- **Data-Driven Decision Making**: Establish a data analytics framework that supports ongoing strategic decisions in marketing and risk management.

Data Mining Goals

Technical Objectives

1. **Develop a Predictive Model**: Build a machine learning model, specifically a Random Forest classifier, which can accurately predict customer eligibility based on a diverse set of variables.
2. **Data Integration and Preprocessing**: Collect and preprocess data from various sources, ensuring data quality, consistency, and suitability for modeling.
3. **Feature Selection and Engineering**: Identify and engineer key features that significantly impact eligibility predictions to enhance model performance.
4. **Model Evaluation and Validation**: Rigorously test the model using appropriate evaluation metrics to ensure reliability and robustness.
5. **System Integration**: Deploy the model within LangaSat's existing CRM infrastructure using Shiny Server for real-time accessibility.
6. **Ethical Data Handling**: Implement data privacy and security measures to protect sensitive customer information throughout the data mining process.

Success Criteria from a Technical Perspective

- **Model Performance Metrics**: Achieve a minimum of 99% accuracy, 99% precision, and 99% recall on the test dataset.
- **Data Quality Standards**: Ensure that 100% of the data used for modeling is cleaned, consistent, and free of critical errors.
- **Feature Impact**: Successfully identify and incorporate at least five additional significant variables beyond annual salary that improve model predictions.
- **System Responsiveness**: The deployed model should provide eligibility assessments within 2 seconds per request to ensure real-time usability.
- **Integration Success**: Seamless integration with the CRM system, with no major technical issues during deployment.
- **Compliance and Security**: Full compliance with data protection regulations and internal security policies, with no breaches or violations.

Data Exploration

Initial Data Collection

The initial dataset was sourced from LangaSat's customer data records, containing attributes pertinent to financial stability, socio-economic indicators, and demographic characteristics of potential customers. This dataset served as the foundation for developing a predictive model to determine customer eligibility for LangaSat's satellite internet services. Key variables in the dataset included Annual Salary, Years of Residence, Gross Year-To-Date Earnings, Education, Occupation, Marital Status, and Household Size, which were anticipated to influence customer eligibility.

Dataset Description

The dataset comprised approximately 19,000 customer records, with each entry representing a unique individual. The columns were a mix of categorical, numerical, and text fields, including:

- **Financial Variables:** Annual Salary, Gross Pay Last Paycheck, Gross Year-To-Date Earnings

- **Demographic Variables:** Year of Birth, Marital Status, Household Size, Years of Residence

- **Geographic and Contact Information:** Street Address, Postal Code, City, State/Province, Country
- **Other Personal Information:** Education Level, Occupation, Phone Number, Email
  The presence of both categorical (e.g., Education, Marital Status) and numerical variables (e.g., Annual Salary, Years of Residence) provided a broad perspective for predicting eligibility based on diverse factors.

Data Quality Assessment

A quality check was conducted to ensure the data's reliability for modeling. This assessment focused on identifying issues that could hinder model performance, such as missing values, duplicate entries, and inconsistent formats.

- **Completeness:** Certain fields, notably Phone Number and Email, contained missing values, due to incomplete records during data collection. Missing values were flagged for further handling during data cleaning.

- **Accuracy:** Numeric fields such as Annual Salary and Years of Residence were verified to be within reasonable ranges based on typical customer demographics. Extreme outliers were identified, which could impact model performance if not addressed.

- **Consistency:** Inconsistent entries were observed in categorical fields like Education and Occupation, where entries such as "Bachelor" and "BSc" referred to the same education level but were entered differently. These inconsistencies were standardized in the cleaning phase to ensure uniformity.

Preliminary Insights

Initial data exploration revealed patterns that hinted at the significance of certain variables in predicting eligibility. For instance, preliminary analysis suggested a positive correlation between Years of Residence and Annual Salary with eligibility likelihood, implying that stable, higher-income customers were more likely to qualify. This insight guided the feature selection process and informed the model's focus on stability and financial status as key predictors.

Data Quality Assessment

Data Completeness

An assessment of data completeness was conducted to ensure all necessary fields had valid values. While core financial variables like Annual Salary and Gross Year-

To-Date Earnings were largely complete, auxiliary fields such as Phone Number and Email had higher rates of missing values. These missing values were either imputed or omitted based on their significance to the predictive model.

Data Accuracy

Data accuracy checks were performed by examining the distributions of numerical variables to identify anomalies. For example, abnormally high values in Annual Salary suggested data entry errors or outliers that needed further scrutiny. Verification against known benchmarks or typical industry ranges was applied to flag potentially erroneous entries.

Data Consistency

Inconsistencies in categorical data were addressed by standardising entries across fields. For example, variations in Education levels, such as "Masters" vs. "M.Sc.", were consolidated. This ensured that all categories were uniformly represented, avoiding ambiguity in model training.

Identified Issues and Challenges

The primary challenges identified included:

- **Handling Missing Values:** Essential for fields like Education and Occupation, which had sporadic missing entries.
- **Outliers in Financial Data:** Extremely high or low salaries were potential outliers that could skew the model.
- **Inconsistent Categorical Data:** Standardisation was required for categorical entries to ensure consistent model input.

Data Preparation

Data Cleaning

Handling Missing Values

Missing values were managed based on the significance of each variable to the model. For example:

- **Imputation:** For numeric fields like Annual Salary, missing values were imputed using median values to maintain data integrity without biasing the data. Categorical fields were filled with the most frequent category where appropriate.

- **Omission:** Non-essential fields with high rates of missing values, such as Phone Number, were omitted to streamline the dataset.

Addressing Outliers

Outliers in financial fields (e.g., Annual Salary and Gross Year-To-Date Earnings) were addressed to prevent skewed predictions. A combination of capping and filtering was applied:

- **Capping:** Extreme values were capped within a certain percentile range to maintain a reasonable distribution.

- **Filtering:** Data entries with values outside of plausible ranges were flagged and either corrected or removed.

Resolving Inconsistencies

Categorical fields like Education and Marital Status required standardization. For example:

- **Consolidation:** Variants like "Single" and "Unmarried" were consolidated into a single category.

- **Case Standardization:** Entries were converted to a uniform case (e.g., "Bachelor" vs. "bachelor") to avoid duplicate categories.

Data Transformation

Data Normalization/Standardization

Numerical fields like Annual Salary and Gross Year-To-Date Earnings were standardized to ensure all numeric variables contributed equally to the model, preventing any single attribute from disproportionately influencing predictions due to scale.

Variable Selection/Elimination

Redundant fields, such as Middle Initial and Street Address, were eliminated, as they had minimal relevance to eligibility prediction. This reduced data dimensionality and streamlined model training.

Feature selection

Modeling

Modeling Techniques

Selected Algorithms

Various algorithms were evaluated to identify the best fit for eligibility prediction. Ultimately, Random Forest was selected due to its robustness, interpretability, and capacity to handle both categorical and continuous variables.

Model Parameters

The key parameters of the Random Forest model were fine-tuned for optimal performance:

- **Number of Trees:** Set to a range that balances accuracy with computational efficiency.

- **Max Depth and Min Samples Split:** Configured to prevent overfitting while capturing complex relationships.

Training Process

To evaluate model performance, the dataset was divided into training and test sets (1split). Cross-validation was employed to ensure that the model generalized well across different data subsets.

Model Building

Model Implementation

The Random Forest classifier was implemented using R's random Forest package. Key steps included data preprocessing, model training, and validation, with each stage documented for reproducibility.

Parameter Tuning

Hyperparameters were optimized using grid search to test different combinations and identify the configuration that yielded the best performance based on accuracy, precision, and recall.

Model Iterations

Several iterations of model training and evaluation were conducted to improve performance metrics. Each iteration provided insights into feature importance, which informed decisions about further refinement or elimination of variables.

Evaluation

Model Performance

Evaluation Metrics

The model's performance was assessed using a range of metrics to ensure accuracy and reliability. Key metrics included:

- **Accuracy**: The proportion of correct predictions out of all predictions made, demonstrating the model's overall effectiveness.

- **Precision**: The ratio of true positive predictions to total positive predictions indicates how well the model avoids false positives.
- **Recall (Sensitivity)**: The ratio of true positive predictions to all actual positive cases, showing the model's effectiveness in identifying eligible customers.

- **F1 Score**: A harmonic mean of precision and recall, providing a single score that balances these two aspects, especially useful when there is a class imbalance.

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: Measures the ability of the model to distinguish between eligible and ineligible customers. An AUC close to 1 indicates excellent model performance.

Results Analysis

The Random Forest model achieved high performance across all metrics, with an accuracy rate close to 99.99%. This high accuracy indicates the model's strong ability to correctly classify eligible and ineligible customers based on multiple variables, moving beyond the initial salary-based approach. Precision and recall were balanced, minimizing false positives and false negatives, and ensuring that eligible customers were accurately identified and ineligible customers were correctly filtered out. This performance aligns well with the business requirements, as it reduces the risk of extending services to high-risk individuals while also ensuring that eligible customers are not overlooked.

Model Comparison

During the development process, multiple algorithms were tested, including logistic regression, decision trees, and k-nearest neighbors. However, Random Forest outperformed the other models in terms of accuracy, precision, and recall, as well as its robustness to overfitting and interpretability. Clustering for example provided some insights into feature importance but did not achieve the same level of predictive accuracy. Ultimately, Random Forest was selected as the final model for its superior performance in balancing accuracy with computational efficiency, making it well-suited for deployment.

Business Value Assessment

Achievement of Business Objectives

The model successfully addresses LangaSat's business objective of creating a reliable, automated eligibility assessment system. By incorporating variables beyond annual salary, the model provides a more comprehensive evaluation of credit risk, allowing LangaSat to make informed decisions on customer eligibility. This shift from a single-variable assessment to a multi-variable model significantly enhances the company's ability to identify and approve customers who represent lower credit risks, thereby aligning with the company's goal of minimizing service-related financial risks.

ROI Analysis

The model is projected to reduce manual evaluation costs and processing time, allowing customer service representatives to focus on high-value tasks rather than spending time on eligibility assessments. The automation provided by the Customer Eligibility System is expected to reduce operational costs related to human resource allocation for manual assessments and improve overall productivity. Additionally, improved accuracy in customer assessment is likely to decrease default rates, further enhancing profitability and providing a tangible return on investment.

Business Impact

Implementing the model within the eligibility assessment process positions LangaSat as a data-driven organization that leverages advanced analytics for decision-making. The model enables LangaSat to streamline its customer evaluation process, leading to faster service delivery, improved customer satisfaction, and a more efficient allocation of company resources. The reliable,

consistent outcomes provided by the model also contribute to strengthening customer trust and the company's reputation in the satellite internet market.

Deployment and Maintenance

Deployment Strategy

Implementation Plan

The Customer Eligibility System was deployed on a secure, scalable server using Shiny Server Pro for interactive, web-based access. The implementation included establishing a secure environment, configuring the database for seamless data storage and retrieval, and integrating the model with the customer relationship management (CRM) system. An initial pilot phase was conducted to test the model in a live environment, with regular monitoring and adjustments made to address any issues.

Integration Requirements

The system was integrated with LangaSat's CRM, allowing customer service representatives to access eligibility predictions directly within their existing workflows. This integration required configuring API endpoints for data flow between the CRM and the model, as well as setting up secure access protocols to maintain data integrity. Data inputs, such as customer demographic and financial information, are pulled from the CRM, processed by the model, and then updated back into the CRM for use in customer interactions.

Monitoring Approach

A monitoring framework was established to continuously track the system's performance and model accuracy post-deployment. Key metrics, such as response

time, prediction accuracy, and system uptime, are monitored to ensure the model maintains its reliability over time. Alerts were configured to notify the technical team of any anomalies or performance drops, enabling proactive maintenance and adjustments when necessary. Additionally, regular model validation checks are conducted to assess for data drift or other changes that may affect the model's performance.

Maintenance Plan

Update Procedures

To ensure the model's accuracy remains high, periodic updates are scheduled based on new data or changing business requirements. The update process involves re-training the model on recent customer data, recalibrating parameters if needed, and redeploying the updated model. This process is critical to adapting to shifts in customer profiles or market conditions, which may influence eligibility criteria over time.

Performance Monitoring

A dedicated performance monitoring system tracks key indicators such as accuracy, precision, and system latency. Regular checks are conducted to identify any issues or trends that may impact the model's reliability. These insights help the technical team determine if updates or modifications are necessary to maintain optimal performance.

Technical Support

Technical support is provided through a tiered support framework to address any issues that may arise. Level 1 support handles basic queries related to system usage, while Level 2 support addresses technical issues such as data flow and server performance. Level 3 support involves expert intervention for complex issues like model performance degradation or system failures. This structure ensures that any disruptions to the eligibility system are promptly resolved, minimizing downtime and maintaining service continuity.

Ethical Considerations

Privacy Concerns

The project involved handling sensitive customer information, including personal details and financial data. Data privacy was a top priority, with stringent protocols implemented to ensure customer data is protected at all stages. The system was designed with built-in data encryption for both storage and transmission, safeguarding against potential data breaches. Access to customer data was restricted to authorized personnel only, following strict access control policies.

Bias Assessment

One of the challenges in building the model was ensuring that it remained fair and unbiased, particularly regarding variables that may indirectly relate to protected characteristics. Bias detection techniques were applied to ensure that the model did not inadvertently favor or disadvantage certain customer groups. Variables that could introduce bias, such as geographic location, were carefully reviewed and adjusted as needed to promote fairness.

Mitigation Strategies

To address ethical concerns, several mitigation strategies were incorporated. For instance, sensitive variables were excluded from the model to prevent discriminatory predictions. Furthermore, the model's decision-making process was made as transparent as possible, with key variables and their contributions to the eligibility prediction highlighted. This transparency enables LangaSat to explain eligibility decisions to customers if needed, fostering trust and accountability.

Compliance Considerations

Compliance with data protection regulations was a fundamental aspect of the project. The system was designed in line with data security standards, ensuring compliance with industry best practices. Regular audits are conducted to ensure adherence to these standards, and the technical team receives ongoing training in ethical data handling and privacy protection. By embedding ethical practices into the model's lifecycle, LangaSat upholds its commitment to responsible data use.

Project Review

Lessons Learned

Successes

- **Improved Efficiency**: The Customer Eligibility System successfully automated the eligibility evaluation process, which was previously manual and time-consuming. This efficiency gain allows LangaSat to evaluate more customers in less time while ensuring consistency and accuracy.

- **High Model Accuracy**: The Random Forest model achieved a 99.99% accuracy rate, surpassing initial expectations. This success reflects the value of using advanced machine learning techniques to make data-driven decisions, providing LangaSat with a robust tool to minimize financial risk.

- **Seamless CRM Integration**: Integrating the model with LangaSat's CRM system streamlined access to eligibility assessments, allowing customer service teams to use it directly within their workflows. This seamless integration reduced friction in operational processes and enhanced service delivery.

Challenges

- **Data Quality Issues**: During data preparation, inconsistencies, missing values, and outliers were identified. Cleaning and transforming the data took longer than anticipated, highlighting the importance of thorough data quality assessment and preparation in similar projects.

- **Balancing Model Complexity and Interpretability**: While Random Forest performed exceptionally well, there were challenges in making its predictions fully interpretable for business users. Balancing technical accuracy with interpretability required iterative adjustments to the model and reporting, especially for business stakeholders.

- **Ethical and Privacy Concerns**: Handling sensitive customer data raised concerns around privacy and bias, requiring additional measures for ethical data processing. These considerations involved extra steps for compliance and transparency, underscoring the importance of ethical considerations in machine learning applications.

Areas for Improvement

- **Streamlined Data Preparation**: Future projects could benefit from establishing a more robust data preparation framework upfront, which would help expedite the cleaning and transformation process.

- **Enhanced Model Interpretability**: Consider using interpretable machine learning models, such as decision trees or SHAP (SHapley Additive exPlanations) values for feature importance in complex models. This could help communicate model decisions to business stakeholders more effectively.
- **Proactive Ethical Audits**: Introducing regular, proactive audits for bias and ethical concerns would help ensure the model remains fair and compliant with privacy standards, especially as new data or variables are introduced.

Learning Experience

Knowledge Gained

This project provided valuable insights into the entire lifecycle of a machine learning application, from data preparation and modeling to deployment and ethical considerations. Key knowledge areas included data wrangling techniques, feature engineering, model evaluation, and the practicalities of deploying a model in a business environment.

Skill Development

The team enhanced their skills in data science and machine learning, particularly in the use of Random Forest and model evaluation metrics. Additionally, technical skills in R, Shiny Server, and database integration were strengthened. The project

also fostered important soft skills, such as teamwork, problem-solving, and adaptability in handling unexpected challenges.

Future Applications

The skills and knowledge gained from this project apply to future data science projects, especially those involving predictive modeling and CRM integration. The project has laid a solid foundation for using machine learning to address business challenges, and the experience gained will support future applications in customer segmentation, recommendation systems, and other predictive analytics use cases.

Conclusion and Recommendations

Project Summary

The Customer Eligibility System project aimed to revolutionize LangaSat's eligibility assessment process by developing a machine learning model that evaluates customers based on multiple financial and demographic variables, rather than relying solely on annual salary. Using a Random Forest classifier, the system achieved an impressive accuracy rate of 99.99%, providing fast, reliable, and consistent eligibility assessments. The successful integration of this model with the CRM system has streamlined LangaSat's operations and improved customer service delivery.

Key Findings

- **High Model Performance**: The model's accuracy and reliability underscore the potential of machine learning to support data-driven decision-making in customer eligibility assessments.

- **Ethical Data Handling**: Ensuring ethical and responsible data handling practices is critical, particularly when sensitive customer data is involved. Privacy protection and bias mitigation were essential components of this project.

- **Operational Efficiency Gains**: Automation of the eligibility assessment process has led to reduced manual workload, quicker service delivery, and enhanced operational efficiency.

Future Recommendations

1. **Regular Model Updates**: To ensure continued accuracy, the model should be re-evaluated and retrained regularly with updated customer data, as customer demographics and financial profiles may shift over time.

2. **Enhanced Interpretability**: Implement techniques to improve model transparency for business users, such as using SHAP values or creating simplified visualizations of key influencing factors in eligibility decisions.

3. **Expansion to Other Services**: The success of this project opens up the possibility of applying similar predictive models to other LangaSat services, such as customer retention strategies, upselling, and personalized service recommendations.

4. **Ethical Framework Expansion**: To further strengthen trust with customers, consider developing a formalized ethical framework for all data-driven projects, which includes regular audits, transparency reports, and bias checks to ensure responsible data use.

References

- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium.
- Schröer, C., Probst, M., & Bader, P. (2021). The CRISP-DM model: The new standard for data mining. International Journal of Data Science and Analytics.
- Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
- R Documentation. (n.d.). Retrieved from https://www.rdocumentation.org/
- Shiny Server Documentation. (n.d.). Retrieved from https://docs.rstudio.com/shinyapps.io/

Appendices

Appendix A: Technical Documentation (GitHub link)

Code Repository Link

- The full project code, including data preprocessing scripts, modeling, evaluation, and deployment code, is available in the repository link:

Data Dictionary

- The data dictionary provides a comprehensive overview of all dataset fields, including descriptions of each variable, data types, ranges, and any transformations applied during preprocessing.
- Key variables include:
    - Annual_Salary: Customer's annual income in local currency.
    - Years_of_Residence: Number of years the customer has lived at their current residence.
    - Household_Size: Number of individuals in the customer's household.
    - Marital_Status: Customer's marital status (single, married, etc.).

Model Documentation

- **Model Selection**: Random Forest was chosen due to its robustness, interpretability, and accuracy in handling mixed data types.
- **Hyperparameters**: Key parameters included the number of trees (ntree) set to 100, and max depth set to 10 to balance performance and interpretability.
- **Feature Importance**: The Random Forest algorithm generated importance scores for each feature, highlighting variables such as Annual_Salary and Years_of_Residence as significant predictors.
- **Evaluation Metrics**: Detailed breakdown of accuracy, precision, recall, F1-score, and confusion matrix results.
- **Deployment Method**: The model was deployed using R Shiny for seamless integration with LangaSat's CRM system.
- For further details, refer to the full model documentation in the project repository.