BIN 381

Project Milestone 5

Members

Jo-Anne van der Wath (577394)
Henry Roux (577440)
Armandre Erasmus (577311)
Chaleigh Storm (577716)

Table of Contents

Tables of Figure

Table of Tables

Introduction

In this milestone, we focus on evaluating the results obtained from our modelling efforts and reviewing the entire data mining process from Phase 1 through 4 to ensure that all steps were properly executed. The objective is to assess how well the models meet the business goals and success criteria outlined at the project's inception and to determine the next steps for deployment or further refinement.

Evaluation of Results

Business Goals Recap

To effectively evaluate our models, it's crucial to revisit the business goals established at the project's inception. These goals provide the benchmark against which we measure the success of our modelling efforts.

Primary Objective:

- Develop a classification model that considers more attributes than just salary to accurately determine customer eligibility for LangaSat's satellite internet services.

The core aim is to enhance the current eligibility assessment, which relies solely on annual salary, by incorporating additional customer attributes. This should result in a more nuanced and accurate determination of eligibility, allowing LangaSat to expand its customer base while managing credit risk effectively.

Secondary Objectives:

1. Increase the precision of decisions regarding customer eligibility.
    - Explanation: Precision in this context refers to the model's ability to correctly identify customers who are truly eligible out of all customers predicted as eligible. A higher precision reduces the number of ineligible customers incorrectly classified as eligible, thereby minimizing potential credit risks.
2. Reduce the number of false positives and false negatives to ensure eligible customers are not mistakenly rejected and ineligible customers are not mistakenly accepted.
    - Explanation: False positives (Type I errors) occur when ineligible customers are incorrectly classified as eligible, leading to potential financial loss. False negatives (Type II errors) happen when eligible customers are incorrectly rejected, resulting in missed business opportunities. Reducing both is essential for optimal operational efficiency.
3. Identify key variables (e.g., occupation, years of residence, education level) that significantly impact customer eligibility.
    - Explanation: Understanding which variables most influence eligibility decisions help tailor marketing strategies and risk assessments, which will assist stakeholders in determining a better baseline for eligibility and understanding the key variables will assist in finding the right target market for the services.

4. Present results using visualizations to aid decision-makers.
    - Explanation: Visual representations of data and model outputs enhance comprehension among stakeholders, facilitating better-informed decisions and fostering trust in the model's recommendations.
5. Increase the number of customers eligible for the services.
    - Explanation: By refining the eligibility criteria through a more comprehensive model, LangaSat aims to expand its customer base without compromising on credit risk, thereby increasing revenue potential. The aim is to create a model which classifies more customers as eligible compared to the baseline model which is based on the salary of R50 000, without generating false positives that increases the credit risk and false negative that could decrease the amount of customers and revenue.
6. Improve model performance in terms of accuracy, precision, and recall compared to the baseline model based solely on income.
    - Explanation: The baseline model uses annual salary as the sole predictor of eligibility. The new model should outperform this baseline by effectively utilizing additional variables, resulting in higher predictive performance across key metrics.

Success Criteria

To determine if the objectives have been met, we established specific success criteria:

1. The model accurately classifies customers as eligible, or ineligible (Not eligible) based on their credit risk.
    - Target: High overall accuracy in predictions, indicating reliable performance.
2. Model performance exceeds acceptable performance metrics (accuracy of at least 85%).
    - Target: A minimum accuracy threshold of 85% is set to ensure the model is significantly better than random guessing and provides tangible business value. An accuracy lower than 85% is insufficient and will more likely result in inaccurate predictions (False eligibility and false non-eligibility).
3. There is an increase in the number of customers that are eligible for services.
    - Target: The number of customers that are eligible for services based on the baseline model is compared to the number of customers that are eligible compared to the new model. If the number of the new model is higher, then the target has been reached and the criteria met.
4. Stakeholders can identify which variables contribute most to credit risk and customer eligibility.
    - Target: The model should provide insights into feature importance, allowing stakeholders to understand and potentially act on the factors influencing eligibility.

5. The model should be built on several attributes and not solely focus on the annual salary.
    - Target: The annual salary should not be used in training the model, however salary groups can be engineered and used. This will force the model to consider other attributes for eligibility.
6. The model adheres to ethical guidelines, avoiding biases and discriminatory classifications.
    - Target: The customer data should be kept private and not shared in any way that is not allowed by the customer and meet the standards set out by the POPI Act of 2013. The model must not discriminate against any group and should comply with all relevant laws and ethical standards, ensuring fairness and compliance.

Model Performance

Table 1: Model Performance Metrics

Metric	Accuracy	Precision	Recall	F1-Score
Logistic Regression	94.80%	92.50%	92.60%	92.60%
Decision Tree	96.50%	92.30%	97.50%	94.80%
Random Forest	97.00%	98.42%	93.07%	95.67%

- Accuracy: Measures the proportion of total correct predictions (both true positives and true negatives) among all predictions.
- Precision: Indicates the proportion of true positives among all positive predictions, reflecting the model's ability to avoid false positives.
- Recall: Reflects the proportion of true positives identified out of all actual positives, showing the model's effectiveness in capturing eligible customers.
- F1-Score: The harmonic means of precision and recall, providing a balance between the two.

Evaluation Against Business Goals

1. Accuracy and Performance

- All models exceeded the 85% accuracy threshold:
    - Logistic Regression: 94.8%
    - Decision Tree: 96.5%
    - Random Forest: 97.0%
- Analysis:
    - Random Forest had the highest accuracy, demonstrating superior performance and validating the inclusion of additional variables.
    - Decision Tree also performed well, showing the effectiveness of tree-based methods.

- Logistic Regression surpassed the threshold, confirming its validity as a predictive tool.

2. Precision and Recall

- Precision:
  - Random Forest: Highest precision at 98.42%, meaning it correctly identified eligible customers and minimized false positives.
  - Decision Tree and Logistic Regression: Precision around 92%, indicating more false positives compared to Random Forest.
- Recall:
  - Decision Tree: Highest recall at 97.5%, successfully identifying most eligible customers and reducing missed opportunities.
  - Random Forest: Recall of 93.07%, although it's not as high as the Decision Tree's recall, it's still relatively high.
  - Logistic Regression: Recall of 92.6%, least effective in capturing all eligible customers.

3. Reducing False Positives and Negatives

Table 2: False Positives and Negatives for Models

Model	False Positive	False Negative
Logistic Regression	985	1004
Decision Tree	330	1053
Random Forest	963	195

- Random Forest:
  - Lowest number of false negatives (roughly 80% lower than the other models). Minimized false positives, reducing financial risk from ineligible customers.
- Decision Tree:
  - Lowest number of false positives (roughly 33% lower than the other models). Minimized false negatives, maximizing revenue by not missing eligible customers.
- Balance:
  - Decision Tree is ideal when minimizing false positives is crucial.
  - Random Forest is beneficial when it's important not to miss any eligible customers.

4. Incorporation of Additional Variables

- Variables Used in Building the Models:
    - Title, Department, Gross Pay Last Paycheck, Gross Year to Date, Gross Year to FRS Contribution, Age, Marital Status, Country, Education, Occupation, Household Size, Years of Residence.
- Feature Importance:
    - Logistic Regression: The feature importance cannot be determined as easily using the logistic regression model.
    - Decision Tree: The feature importance can be derived from the decision tree model based on the branches that are generated. According to the tree, 'Gross Pay Last Paycheck' is the most important feature, followed by Salary Group High and Salary Group Low.
    - Random Forest: The random forest model generates feature importance plots. The feature importance according to mean decrease in accuracy states that Gross Pay Last Paycheck, Salary Group Low and Salary Group Medium is most important. If importance is measured according to branch impurity rates, then according to the random forest model, Gross Pay Last Paycheck, Gross Year to Date and Gross Year to FRS Contribution are the most important features.

5. Increasing Eligible Customers

```
## Calculate % eligible of baseline

```{r}
numEligibleOriginal <- sum(customers$Eligible == 1, na.rm = TRUE)
totalCustomersOriginal <- length(customers$Eligible)
eligiblePercantageOriginal <- (numEligibleOriginal / totalCustomersOriginal) * 100
cat("Percentage of Eligible Customers in baseline model: ", round(eligiblePercantageOriginal, 2))
```

Percentage of Eligible Customers in baseline model: 64.43
```

Figure 1 Baseline Eligibility Rate

The baseline model has an eligibility rate of 64.43%. This means that 64.43% of their current customers are eligible for the satellite services. This is based on the annual salary of R50 000.

```
## Calculate Eligibility Rates - Logistic Regression

```{r}
# Assuming `predictions` is a vector of 1s (eligible) and 0s (not eligible) from your model
# For example: predictions <- predict(model, newdata, type = "response") > 0.5

# Count eligible customers
num_eligible_customers <- sum(logisticRegressionY_pred == 1)

# Total number of customers
total_customers <- length(logisticRegressionY_pred)

# Calculate the eligibility percentage
eligibility_percentage <- (num_eligible_customers / total_customers) * 100

cat("Percentage of eligible customers:", round(eligibility_percentage, 2), "%\n")
```

Percentage of eligible customers: 64.48 %
```

Figure 2 Logistic Regression Eligibility Rate

The logistic regression model has an eligibility rate of 64.48%. This means that 64.48% of customers are eligible for the LangaSat services, according to this model. There has only been an increase of 0.05 in the number of customers.

```
## Calculate Eligibility Rates - Decision Tree

```{r}
# Assuming `predictions` is a vector of 1s (eligible) and 0s (not eligible) from your model
# For example: predictions <- predict(model, newdata, type = "response") > 0.5

# Count eligible customers
num_eligible_customers <- sum(decisionTreePredictions == 1)

# Total number of customers
total_customers <- length(decisionTreePredictions)

# Calculate the eligibility percentage
eligibility_percentage <- (num_eligible_customers / total_customers) * 100

cat("Percentage of eligible customers:", round(eligibility_percentage, 2), "%\n")
```

Percentage of eligible customers: 66.32 %
```

Figure 3 Decision Tree Eligibility Rate

The decision tree model has a customer eligibility rate of 66.32%. There is a visible increase of 1.89% in the number of customers that are considered as eligible for the services.

```
## Calculate Eligibility Rates - Random Forest

```{r}
# Assuming `predictions` is a vector of 1s (eligible) and 0s (not eligible) from your model
# For example: predictions <- predict(model, newdata, type = "response") > 0.5

# Count eligible customers
num_eligible_customers <- sum(randomForest_predictions == 1)

# Total number of customers
total_customers <- length(randomForest_predictions)

# Calculate the eligibility percentage
eligibility_percentage <- (num_eligible_customers / total_customers) * 100

cat("Percentage of eligible customers:", round(eligibility_percentage, 2), "%\n")
```

Percentage of eligible customers: 66.44 %
```

Figure 4 Random Forest Eligibility Rate

The random forest model has an eligibility rate of 66.44%. This is the highest number of eligible customers for all the models. From the baseline model, there is an increase of 2.01% in the number of customers that are labelled as eligible for the services.

- Outcome:
    - The random forest model showed the highest increase in number of eligible customers.
    - Benefit: Expanded the customer base, tapping into previously overlooked market segments.

6. Visualization and Interpretability

- Decision Tree Model:
    - Provided clear visual decision pathways.
    - Advantage: Easy for stakeholders to understand decision-making processes.
- Visualizations Created:
    - Correlation Matrices: Showed relationships between variables.
    - Feature Importance Plots: Highlighted influential predictors.
- Business Impact:
    - Enhances transparency and trust.
    - Facilitates effective communication between technical and business teams.

Approval of Models Meeting Business Success Criteria

Random Forest Model

- Assessments:
    - Performance: Highest accuracy (97.0%) and precision (98.42%). The random forest has the best performance in comparison with the other models. The model accurately makes predictions to classify customers as eligible or not eligible.
    - Metrics: Minimizes false positives, reducing credit risk, and expands the customer base. Generates the lowest number of false negatives, therefore approving more eligible customers.
    - Features: Feature importance is calculated by the model and is clearly represented in importance plots for ease of use. Furthermore, the model was trained using many variables and was not built solely based on annual salary. Provides valuable information for strategic decisions.
    - Eligibility: The number of eligible customers increases with 2.01% when this model is implemented, giving the greatest increase in eligible customers among the models.
    - Visualisation: Feature importance plots are easily generated using the random forest model.
    - Approval: This model meets all the requirements as set by the success criteria. Furthermore, the overall performance of this model is the best of the three. This model is approved and will be implemented and deployed in the following milestones.
- Considerations:
    - Resource Requirements: May need more computational resources, but benefits justify the investment.

Decision Tree Model

- Assessment:
    - Although the decision tree model also has very high-performance metrics, it is slightly worse than random forest. Furthermore, the increase in customers that are eligible are not as many as with the random forest model. Although the most importance features can be derived from the decision tree, it is not as complete as the importance plots generated by the random forest model.
    - Approval: This model will not be approved or deployed.
- Considerations:
    - Simplicity and Interpretability: Easy to implement and understand.
    - Visualization: Clear decision rules beneficial for stakeholders.
    - Lower Precision: Higher rate of false positives may increase credit risk.

Logistic Regression Model

- Assessment:
    - Performance: This model has a lower performance than the other models but exceeds minimum requirements. Furthermore, it does not provide the importance of features. The increase in the number of eligible customers is so little it will not be considered an increase.
    - Approval: This model will be rejected.
- Considerations:
    - Lower Predictive Power: Less effective with complex, nonlinear relationships.

Conclusion:

After assessing all three of these models according to the success criteria, it has been concluded that all three models can be used for the project seeing as they meet the criteria and objectives. This project however only seeks to make use of one model. Therefore, the Random Forest Model will be chosen as the best and most significant model and will be deployed.

Reviewing the Process

Phase 1 Business Understanding

1. Define Business Problem
    a. Summary: The business problem clearly defines the need for a classification model to determine eligibility, taking more attributes into account then only annual salary as LangaSat is currently doing.
    b. Issues: No obvious issues.
    c. Improvements: In future projects the stakeholders may be asked to provide clarity on their main requirements of the project.
2. Define business Objects
    a. Summary: The business objectives clearly define the objectives as required by the stakeholders. This includes the main objective, which requires the classification model as stated in the business problem. The secondary objectives are also defined, including the need for precision, as well as accuracy. The objectives clearly state the need for feature importance to be determined. Furthermore, the number of eligible customers need to be increased according to the objectives.
    b. Issues: The goals for the % of precision, recall and F1 scores are not clearly defined, and are only required to be 'good'. There is no actual number to which we can compare to clearly classify the goals as being met. The secondary objectives were not explained / elaborated on enough.
    c. Improvements: The secondary objectives were further expanded in this document. In future projects, more clear performance measurement goals will be set.
3. Identify Stakeholders
    a. Summary: The main stakeholders were identified as LangaSat. Secondary stakeholders of the project include the project supervisor, Mr Gift Mudare, as well

as the development team. Stakeholder requirements were established in the first phase.
    b. Issues: The project supervisor was not consulted enough throughout the development process.
    c. Improvements: In future projects, the project supervisor will be consulted with more regularly, and guidance will be delivered from the supervisor to the development team.
4. Define Success Criteria
    a. Summary: The success criteria was thoroughly defined.
    b. Issues: The success criteria may also seem redundant and repetitive.
    c. Improvements: The success criteria were also adjusted by removing repetitive criteria and refining the remaining ones.
5. Data Mining Goals
    a. Summary: The data mining goals include the points that were made in the objectives and success criteria.
    b. Issues: The data mining goals are to some degree redundant.
    c. Improvements: In future projects, the data mining goals may need to be more detailed and clearly defined.
6. Risk Analysis
    a. Summary: The risks associated with the project include data quality problems, biases in the data, overfitting of the model, issues concerning data privacy and time constraints associated with the project. The risk, severity, likelihood, impact and possible mitigation strategies were discussed in this phase.
    b. Issues: No obvious issues were identified in the risk assessment.
    c. Improvements: In future projects, more potential risks may be identified to ensure that all risks were considered and planned for.

Phase 2 Data Understanding

1. Gather Data
    a. Summary: The CustData2.csv file will be the data source for this project.
    b. Issues: No clear issues.
    c. Improvements: None currently.
2. Understand Data
    a. Summary: The overall data in this dataset was examined and each attribute was analysed to understand its datatype and description. This provided an understanding of the dataset as a whole and gave a holistic understanding of the data.
    b. Issues: There were no issues at this stage of the phase.
    c. Improvements: None currently.
3. Data Visualisation
    a. Summary: Initial visualisation was done at this stage. A heat matrix was created to examine the correlation between attributes. Pair plots were created to visualize the relationships between highly correlated attributes. Boxplots were created for all numerical attributes to visualize outliers. Scatterplots and histograms were created to visualize the distribution of attributes. An initial dashboard was also created.

   b. Issues: Minor issues were experienced during the creation of the initial dashboard. The uncleaned data made it difficult to create visualizations, because of the high cardinalities and numerous empty cells within the dataset.
   c. Improvements: In future projects, the data cleaning will be done before the dashboard will be created. This will allow for ordered, clear dashboards.
4. Data Quality
   a. Summary: The quality of the data depended on the amount of data that was useable. Some attributes had high cardinalities and therefore had the more than twice the number of unique values than supposed to.
   b. Issues: Through continuation of the next phases, it became clear that some cleaning steps were missed or done incorrectly.
   c. Improvements: All improvement were made during milestones 3 and 4.
5. Data Aggregation
   a. Summary: During data aggregation, the data was summarized and grouped to provide statistical analysis on specific data. The following aggregation calculations were done.
      i. Sum of Annual Salary by Department Name.
      ii. Average Annual Salary by Title.
      iii. Number of Customers by Education Level.
      iv. Average Gross Year to Date by Age.
      v. Average Household Size by Years in Residence.
      vi. Average Annual Salary by Education Level.
      vii. Average Age by Occupation.
      viii. Number of Customers by Country
   b. Issues: Although the aggregations provide summaries of relevant features, the summaries may not be completely relative in terms of the features that are highly correlated.
   c. Improvements: In future projects, more relevant aggregation may be done on more highly correlated features.

Phase 3 Data Preparation

1. Select Data
   a. Summary: In the CustData2.csv which contains the data, the following attributes were selected to be used in the training of the model: Title, Department Name, Annual Salary, Gross Year Last Paycheck, Gross Year to Date, Gross Year to Date FRS Contribution, Age, Marital Status, Country ID, Education, Occupation, Household Size, Years in Residence and Eligible. Although Annual Salary is not directly used in the model, it is used to create the feature Salary Groups which will also be used in the training of the model. Furthermore, Eligible is the target attribute that is created based on the annual salary of R50 000.
   b. Issues: Originally, an unsupervised model was selected, and as a result the Eligible model was not created yet. It was later added.
   c. Improvements: In the future, the type of learning that will need to be done will need to be considered more carefully to ensure proper data selection is done from the start.

2. Clean Data
    a. Summary: The data cleaning was done quite thoroughly. The following steps were taken during the cleaning of the data:
        i. Keeping relevant columns.
        ii. Cleaning marital status to ensure there were only six types of statuses. There were different versions of the same status that needed to be formatted into the six statuses.
        iii. Empty cells in marital status were replaced by the mode (most common) status.
        iv. There was only a number of records that were almost completely empty. As the records were almost entirely incomplete, the best course of action was to remove the empty records.
        v. Outlier treatment was performed on Annual Salary, Gross Year Last Paycheck, Gross Year to Date and Gross Year to Date FRS Contribution.
        vi. The columns were renamed to ensure that all columns follow the same naming conventions.
    b. Issues: Age was not checked for outliers.
    c. Improvements: In the future, numerical attributes will be examined carefully to ensure that all values are checked for outliers, and these are handled.
3. Transform Data
    a. Summary: After the data cleaning has been completed, data transformation was done. The steps done in data transformation include the following:
        i. The Eligible attribute was created based on the Annual Salary. This is the target variable.
        ii. Customer Age was created based on the Year of Birth.
        iii. The cardinality of categorical attributes were checked and based on that, the following attributes were factored: Marital Status, Education and Occupation.
        iv. Salary Groups were created by binning the Annual Salary. This produced low, medium and high salary groups.
        v. One hot encoding was done on the following attributes: Marital Status, Education, Occupation and Salary Groups.
        vi. Frequency encoding was done on the following attributes: Title, Department Name and Country ID.
        vii. The skewness of the numerical attributes was checked. Accordingly, standardisation was applied to the attributes. Robust scaling was done on Annual Salary, Gross Year Last Paycheck, Gross Year to Date, Gross Year to Date FRS Contribution. Z-score normalisation was done on age.
        viii. The attributes to keep in the dataset were revied, and the following attributes were kept and saved to be used during modelling: Annual_Salary, Gross_Pay_Last_Paycheck, Gross_Year_To_Date, Gross_Year_To_Date_FRS_Contribution, Age, Household_Size, Years_Residence, Marital_Statusdivorced, Marital_Statusmarried, Marital_Statussingle, Marital_Statuswidowed, EducationBach., EducationHS-grad, EducationMasters, OccupationCleric., OccupationExec., OccupationProf., OccupationSales, Salary_GroupLow, Salary_GroupMedium, Salary_GroupHigh, Salary_GroupVery_High,

Frequency_Title, Frequency_Department, Frequency_Country_ID, and Eligible.

b. Issues: Some issues during the data transformation stage include not balancing the data. When data balancing was applied, the model became overfitted, which resulted in unrealistic outcomes from the model. Furthermore, regularisation was not done to prevent overfitting of the model.
c. Improvements: In future projects, regularisation will be implemented to ensure that model overfitting or underfitting does not occur.

Phase 4 Modelling

1. Select Modelling Techniques
    a. Summary: Three modelling techniques were considered for this project. Logistic regression, decision tree and random forest were the techniques that were chosen based on their classification abilities.
    b. Issues: Initially, a clustering model was considered, which caused a delay in the modelling.
    c. Improvements: In the future, the objectives will be more clearly examined to accurately determine which modelling techniques will be the most appropriate to use.
2. Splitting of Data
    a. Summary: The dataset is split into 80% for training and 20% for testing.
    b. Issues: No issues.
    c. Improvements: None currently.
3. Build Models
    a. Summary: The logistic regression model, decision tree model and random forest model were all created and saved.
    b. Issues: The models do not show any signs of issues.
    c. Improvements: None currently.
4. Assess Models
    a. Summary: The accuracy, precision, recall and F1-score of all three models were calculated and assessed. All three models met the baseline requirement for the desired accuracy of the models.
    b. Issues: No issues were identified during the assessment of the models.
    c. Improvements: None currently.

Determining the next steps

Possible Actions After This Phase

After completing the evaluation and reviewing the entire process, we have identified several potential courses of action. Each option has its own set of advantages and disadvantages, which are elaborated below to aid in making an informed decision.

1. *Deploy the Random Forest Model*

Pros:

- Immediate Improvement in Customer Eligibility Assessments:

  Deploying the Random Forest model will provide instant enhancements to the way LangaSat assesses customer eligibility. With an accuracy of 97.0% and precision of 98.42%, the model significantly outperforms the current salary-only criterion. This means that the company can start making better-informed decisions right away, leading to increased efficiency and effectiveness in customer acquisition.

- Alignment with Business Objectives:

  The model's high accuracy and precision directly align with LangaSat's business objectives of expanding the eligible customer base while maintaining low credit risk. By accurately identifying eligible customers, the company can increase its market share and revenue without incurring additional risk (Breiman, 2001).

Cons:

- Resource Requirements for Deployment and System Integration:

  Implementing the model into the existing operational framework will require resources, both in terms of time and finances. This includes the development of necessary infrastructure, integration with current systems, and ensuring data pipelines are in place for real-time or batch processing. According to Kuhn (2019), successful deployment requires careful planning and allocation of technical resources.

- Staff Training May Be Necessary:

  Employees involved in customer assessment and related processes may need training to understand and effectively use the new system. This includes interpreting model outputs, integrating insights into decision-making, and maintaining the system post-deployment. Training programs will incur additional costs and require time investment from staff.

2. *Further Iterate and Improve the Model*

Pros:

- Potential to Enhance Performance Through Additional Data and Tuning:

  By collecting more data and performing advanced hyperparameter tuning, there is potential to improve the model's performance even further. This could involve addressing issues such as class imbalance using techniques like SMOTE (Chawla, et al., 2002), exploring feature

engineering opportunities, and refining the model through cross-validation and other optimization methods.

- Address Identified Issues:

Iterating on the model allows the team to tackle identified concerns, such as the limited hyperparameter tuning performed in the initial modelling phase, and the exclusion of annual salary potentially omitting important interactions. This can lead to a more robust and generalizable model that may perform better in the long term.

Cons:

- Delays Deployment and Postpones Realization of Benefits:

Further development will delay the implementation of any improvements the current model could offer. This postponement means the company continues to operate without the benefits of the enhanced eligibility assessments, potentially missing out on immediate gains in efficiency and revenue.

- Additional Costs and Resource Allocation:

Additional iterations require more resources, including data scientists' time, computational power, and possibly new data acquisition. This could strain budgets and divert resources from other critical projects. As noted by Hastie, et al. (2009), extensive model tuning can be resource-intensive with diminishing returns after a certain point.

3. *Combine Models for Enhanced Performance*

Pros:

- Balancing Precision and Recall Through Ensemble Methods:

Combining the Random Forest and Decision Tree models could leverage the strengths of both. While Random Forest offers high precision, the Decision Tree excels in recall. An ensemble approach could balance these metrics, potentially capturing more eligible customers while still minimizing false positives. Ensemble methods are known to improve predictive performance by aggregating the predictions of multiple models (Hastie, et al., 2009).

- Capturing Complex Patterns:

Using multiple models may allow for a more nuanced understanding of the data, capturing complex patterns that a single model might miss. This could lead to better overall performance and more accurate eligibility assessments.

Cons:

- Increased Complexity in Model Management:

Managing multiple models increases the complexity of the system. It requires more sophisticated infrastructure to deploy, maintain, and update the models. This complexity can lead to challenges in troubleshooting, version control, and ensuring consistent performance across models.

- Higher Computational Resources Required:

Ensemble models demand more computational power, both during training and inference. This could result in higher operational costs and may require investment in additional hardware or cloud computing resources. According to Breiman (2001), ensemble methods like Random Forest are already computationally intensive, and adding more models exacerbates this issue.

4. *Abandon and Restart the Project*

Pros:

- Opportunity to Reassess the Approach Entirely:

Starting over allows for a fresh perspective on the problem. It provides an opportunity to incorporate new methodologies, data sources, or technologies that were not considered initially. This could potentially lead to a more innovative and effective solution that better aligns with evolving business needs.

Cons:

- Loss of Time and Resources Invested:

Abandoning the project results in a loss of all the effort, time, and resources already expended. The company would forfeit the immediate benefits that could be gained from deploying the current model. Additionally, the sunk cost in terms of personnel hours and financial investment would not yield any return.

- Business Objectives Remain Unmet:

The original business objectives, such as improving customer eligibility assessments and expanding the customer base, would remain unaddressed. This could have negative implications for the company's growth and competitiveness in the market.

Decision on Project Proceeding

After carefully considering the pros and cons of each option, the recommended course of action is to proceed with deploying the Random Forest model while planning for future iterations to refine and enhance the model.

Justification:

- Balance of Immediate and Long-Term Benefits:

Deploying the model now allows LangaSat to begin reaping the immediate benefits of improved customer eligibility assessments, increased efficiency, and potential revenue growth. At the same time, planning for future iterations ensures that the model can be refined and optimized over time, addressing any outstanding issues without delaying current advantages.

- Resource Optimization:

This approach makes efficient use of resources by capitalizing on the work already completed. It avoids the additional costs and delays associated with further immediate iterations or combining models, while still allowing for enhancements in the future.

- Alignment with Business Objectives:

The Random Forest model meets and exceeds all the business success criteria established at the project's inception. Deploying it aligns directly with the company's goals of expanding the eligible customer base and maintaining low credit risk.

- Risk Mitigation:

Delaying deployment to further iterate or combining models introduces risks such as market changes, competitor advancements, or internal shifts in priorities. By deploying now, LangaSat mitigates these risks and positions itself advantageously in the market.

- Flexibility for Future Improvements:

Deploying the model does not preclude future enhancements. The company can monitor the model's performance and plan for iterative improvements based on real-world results and feedback. This allows for a data-driven approach to refinement, ensuring that resources are allocated effectively.

Implementation Plan:

- Deployment Preparation:

Allocate necessary resources for deployment, including system integration and staff training. Develop a detailed plan to integrate the model into existing workflows, ensuring minimal disruption to operations.

- Monitoring and Evaluation:

Establish metrics and processes to monitor the model's performance post-deployment. This includes tracking accuracy, precision, recall, and impact on business KPIs.

- Planning for Iteration:

Schedule future iterations to address identified issues such as hyperparameter tuning and class imbalance. Set timelines and allocate resources accordingly, ensuring continuous improvement.

- Stakeholder Communication:

Keep all relevant stakeholders informed about the deployment and future plans. Gather feedback from users to inform subsequent iterations.

Conclusion

The project successfully developed a predictive model that aligns with LangaSat's business goals of accurately determining customer eligibility using factors beyond annual salary. The Random Forest model is approved for deployment due to its superior performance in accuracy and precision, effectively reducing false positives and improving customer selection (Breiman, 2001).

All four of the previous phases of the CRISP-DM methodology were assessed in detail. Each step completed for each phase was analysed and any issues were identified and explained. Solutions for all the identified issues are provided and plans to improve in future projects were described.

Immediate Business Value

Deploying the Random Forest model will provide **immediate business value** in several ways:

1. **Expansion of Eligible Customer Base**: By considering additional variables such as occupation, years of residence, education level, and household size, the model identifies customers who may have been overlooked under the previous income-only criterion. This broader evaluation increases the pool of eligible customers, potentially leading to higher sales and revenue growth. According to Breiman (2001), ensemble methods like Random Forest are effective in capturing complex patterns in data, which can unveil new customer segments.

2. **Maintenance of Low Credit Risk**: Despite expanding the customer base, the model maintains a low credit risk by accurately distinguishing between eligible and ineligible customers. Its high precision (98.42%) ensures that false positives are minimized, reducing the likelihood of extending services to high-risk individuals (Evidently AI, 2024). This balance is crucial for sustaining profitability and avoiding losses associated with defaults.

3. **Improved Decision-Making**: The model provides insights into which variables most significantly impact eligibility. For instance, features like Gross_Pay_Last_Paycheck and Salary_GroupLow were identified as important predictors. This information enables LangaSat to tailor marketing strategies and focus on customer segments with the highest potential, optimizing resource allocation.

4. **Competitive Advantage**: Implementing advanced analytics distinguishes LangaSat from competitors who may still rely on less sophisticated methods. This technological edge can enhance the company's reputation as an innovative and customer-centric provider.

Continuous Monitoring and Refinement

To ensure the model remains effective and continues to deliver value, continuous monitoring and refinement are essential:

1. **Adaptation to Changing Data Patterns**: Customer behaviors and market conditions are dynamic. Regularly updating the model with new data allows it to learn from recent trends and adjust its predictions accordingly. This adaptability ensures sustained accuracy over time (Hastie, et al., 2009).

2. **Alignment with Business Needs**: As LangaSat's objectives evolve, the model can be recalibrated to reflect new priorities, such as targeting different demographics or adjusting for economic shifts. Continuous refinement ensures the model stays aligned with strategic goals.

3. **Feedback Integration**: Incorporating feedback from customer interactions and sales outcomes helps identify areas where the model may need improvement. This iterative process enhances the model's performance and reliability (Kuhn, 2019).

4. **Compliance and Ethical Considerations**: Ongoing monitoring ensures that the model adheres to ethical guidelines and regulatory requirements, especially concerning data privacy and discrimination avoidance (Roshan, et al., 2024).

5. **Performance Tracking**: Establishing key performance indicators (KPIs) such as accuracy, precision, recall, and the impact on customer acquisition rates allows for quantitative assessment of the model's effectiveness. Tracking these metrics over time facilitates proactive adjustments.

6. **Scalability and Integration**: As the business grows, the model can be scaled to handle larger datasets and integrated with other systems like customer relationship management (CRM) platforms for seamless operations.

Sustaining Effectiveness Over Time

By committing to continuous improvement, LangaSat ensures that the Random Forest model remains a valuable asset:

- **Proactive Risk Management**: Regular assessments can identify potential issues such as data drift or model degradation before they impact performance, allowing for timely interventions.

- **Innovation and Technological Advancement**: Staying informed about advancements in machine learning enables the incorporation of new techniques that could enhance model capabilities.

- **Customer Satisfaction**: A model that adapts to customer needs and market trends contributes to better customer experiences, fostering loyalty and positive brand perception.

- **Strategic Decision Support**: The insights derived from the model support strategic planning and decision-making at higher organizational levels.

Final Thoughts

In summary, deploying the Random Forest model offers immediate benefits by improving customer eligibility assessments and expanding the customer base without increasing credit risk. Continuous monitoring and refinement are critical to maintaining these benefits, ensuring the model adapts to changing conditions and continues to align with LangaSat's business objectives.

By embracing this data-driven approach, LangaSat positions itself for sustained success in a competitive market. The model not only enhances current operations but also provides a foundation for future innovations and growth opportunities. This strategic investment in advanced analytics and continuous improvement reflects a commitment to excellence and customer-centric service delivery.

References

Breiman, L., 2001. Random Forests. In: *Machine Learning*. s.l.:SpringerLink, pp. 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research,* 16(16), pp. 321-357.

Evidently AI, 2024. *Accuracy vs. precision vs. recall in machine learning: what's the difference?*. [Online]
Available at: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall
[Accessed 14 October 2024].

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning*. 2nd ed. s.l.:SpringerLink.

Kuhn, M., 2019. *The caret Package*. [Online]
Available at: https://topepo.github.io/caret/
[Accessed 15 October 2024].

Roshan, K., Bamini, J., Jakhongirov, I. & Silpa, G., 2024. Finding the Factors by Integrating the AI Model and Business Analytics Risk Assessment on Firm Organization. *IEEE Xplore,* pp. 1074-1080.