

BIN381

Project

Group F

**Business
Understanding**

2

3

4

5

6

Problem

The model LangaSat currently uses is based on a salary threshold of R50 000.

**Objectives
& Goals**

- Create a classification model
- Do not solely consider Annual Salary for model creation
- Accurately identifies the eligibility of customers
- Increase the accuracy of decisions
- Reduce false positives and false negatives
- Increase the number of eligible customers
- Find important indicators of eligibility
- Visualisations
- Accuracy over 85%

Business Understanding

2

3

4

5

6

Risks

Assumptions

Constraints

Data Quality
Overfitting
Biased Data
Data Privacy
Missing Deadlines

Complete Data
Necessary Data
Data Cleaning will be sufficient
Relevant Data
Free of Bias
R and Power BI
Currently Annual Salary
Threshold

CRISP-DM
Quality of dataset
Size of the dataset
Development Team
Time Frame
Ethical Guidance

1

Data Understanding

3

4

5

6

Data Overview

Relevant Columns: Title, Department Name, Annual Salary, Gross Pay Last Paycheck, Gross Year to Date, Gross Year to Date FRS Contribution, Year of Birth, Marital Status, Country ID, Education, Occupation, Household Size, Years in Recidance and Eligible.

Data Quality

Missing Data
Duplicates
Outliers
Many versions of the same groups
Variable types
Cardinality



Data Cleaning and Transformation

- Feature engineering
- Remove irrelevant columns
- Recode similar attributes
- Fill empty values
- Remove empty rows
- Outlier Treatment
- Frequency Encoding
- One-hot encoding

1

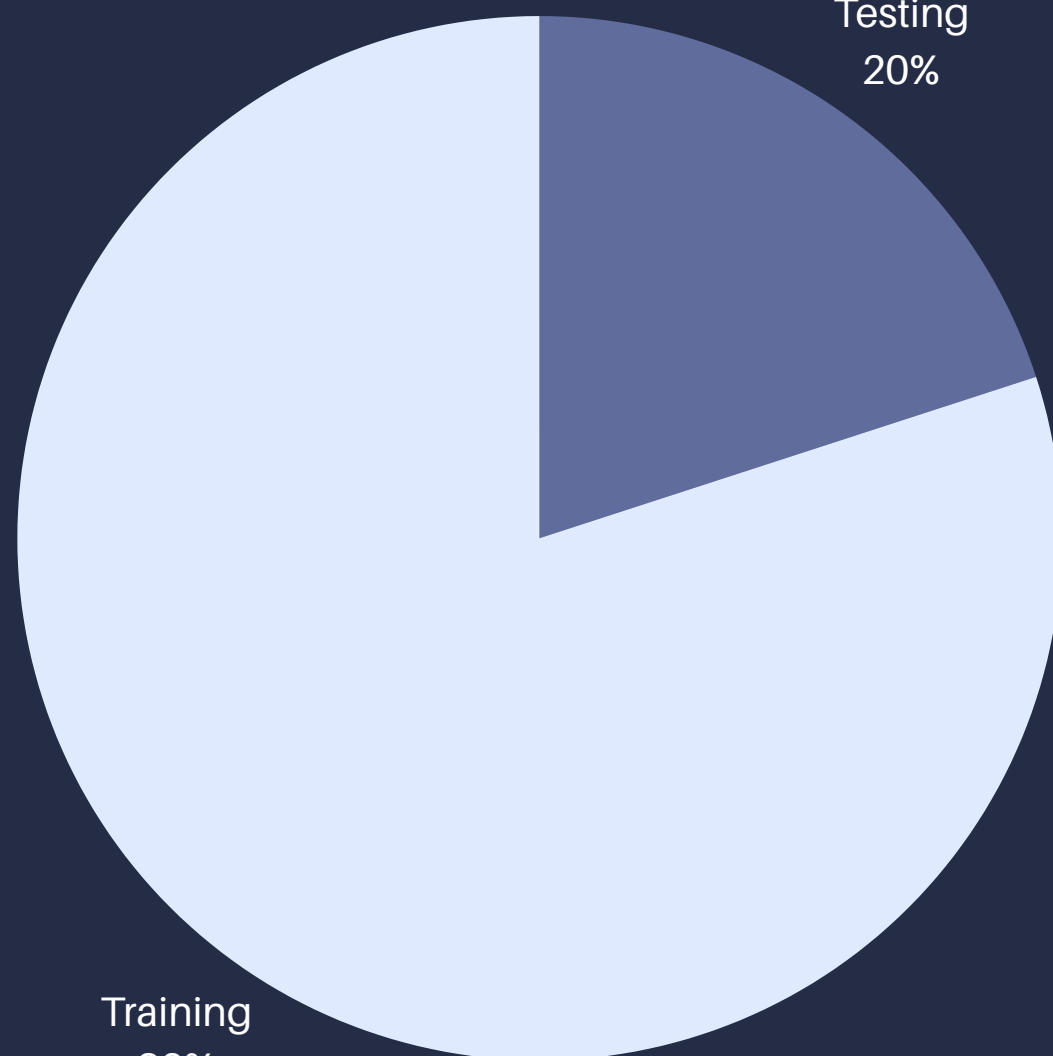
2

3

Modelling

5

6



Training
80%

Testing
20%

Model

Random Forest Model

1

2

3

4

Evaluation

6

Positive

Negative

Positive

12481

368

Negative

1129

24285

Accuracy

96.1%

Precision

97.1%

Recall

91.7%

F1 Score

94.3%

64.4%

Percentage of Eligible
Customers

66.4%

1

2

3

4

5

Deployment

Shiny App

Load a
Record

Bulk
Upload

View All
Records

Model
Metrics

Group
Members

Ethical Considerations

Data Privacy & Confidentiality

Data Minimization & Relevance

Bias & Fairness

Transparency & Interpretability

Data Security

Ethical Use of Predictions

Group Experience

Predictions
Decision Tree
Practical Solutions
Interface
Teamwork
CRISP-DM

Pipeline

Machine Learning

Shiny

Logistic Regression
Data
Preprocessing
Model Outcomes
Random Forest

Server

Goals

Real World

The background features a dark blue field with horizontal, textured brushstrokes in a slightly lighter shade of blue. In the corners, there are light blue geometric shapes: a stepped corner piece in the top-left and bottom-right, and two sets of overlapping circles in the top-right and bottom-left. The word "Conclusion" is centered in a bold, white, sans-serif font.

Conclusion

The background is a dark navy blue. In the center, there is a horizontal, textured brush stroke in a slightly lighter shade of blue. The words "Thank You" are written in a bold, white, rounded sans-serif font across the middle of this brush stroke. To the left of the text, there are two overlapping circles: a larger one in the back and a smaller one in the front, both with white outlines and a slight blue shadow. To the right of the text, there are also two overlapping circles: a larger one in the back and a smaller one in the front, both with white outlines and a slight blue shadow. In the top-left corner, there is a white, stylized, wavy line that looks like a corner bracket or a decorative flourish. In the bottom-right corner, there is a similar white, stylized, wavy line, also acting as a decorative corner element.

Thank You