Visa for Lisa:

Predicting Clients for Personal Loans

Presentation by **Nikita Gaidamachenko**

Table of contents

01 Task

Goals posed for this project

02 Data Cleaning

What information is missing or is misrepresented?

03 Data Exploration

What characteristics do people who accepted loans have in common?

04 Choosing Model

What prediction model is most suitable and why?

05 Model Performance

How accurate is the model in identifying the correct answer?

06 Implications

Implications for clients, marketing and business

01

Task

Task

Description

- 9% of clients offered a loan by Galaxy Bank accept it and become a loan customers.
- Help marketing devise campaigns with better target marketing to increase the success ratio with a minimal budget.

Goals

- Better predict and identify who will accept loans offered to potential loan customers.
- Implement a multi-variable prediction model on a large and complex data set.
- Analyze and evaluate the risks to the business of the implications, assumptions, and decisions in the model.

02

Data Cleaning

Data Cleaning

Errors in Data

- Clients with CC Average set as not owning a Credit Card
- Client with no CC Average set as owning a Credit Card
- Rows with missing or wrong values (such as value set to 2 in binary classification instead of 0 or 1)

Cleaning Data

- Set all clients with any CC Average over 0 to owning a Credit Card
- Set all clients with CC Average of 0 to **not** owning a Credit Card
- Removed rows that contain errors

03

Data Exploration

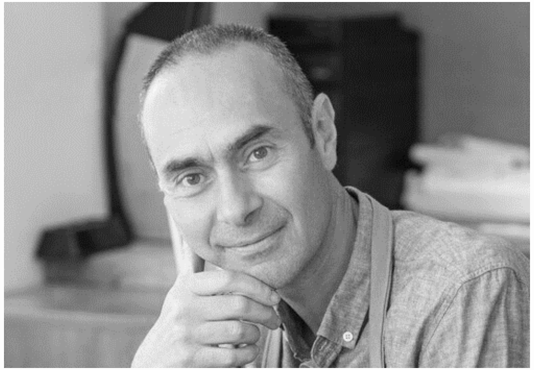
Personal Loan Client

Accepted

Age: 45 years old
Family of: 3-4
Experience: 20+
Education: Post-Secondary+
Income: $144,750
$60,000 (at the lowest)
CC Average: $3,905
Mortgage: $288,000, *if active*
CD Account: 1 in 3 people

Top-5 Criteria

CD Account	5%
Family	12%
CC Average	16%
Education	18%
Income	32%



Declined

Age: 45 years old
Family of: 1-2
Experience: 20+
Education: Secondary+
Income: ≈ $60,000
CC Average: $1,730
Mortgage: $170,860, *if active*
CD Account: 1 in 29 people

Personal Loan Client

1

Age

26-65 years old
Average: 45

2

Family

Clients with family of 3-4 are more likely to accept a loan:

- Family of 1: 22.3%
- Family of 2: 22.0%
- Family of 3: 27.7%
- Family of 4: 28.0%

3

Education

- 20% have first level of education
- 38% have undergraduate or some college education
- 42% have degree or graduate studies

4

Mortgage

- 65% do not have a mortgage
- 35% have an active mortgage

$288,000 – average active mortgage

5

Experience

- 5% entry level (<2 years)
- 8% junior level (3-5)
- 14% mid level (6-10)
- 25% senior level (11-20)
- 48% expert level (20+)

6

Income

- 8.6% less than $100,000
- 46% $100,000-$150,000
- 45.4% over $150,000

$60,000 is the lowest

04

Choosing Model

Alternative Rejected Models

Logistic Regression

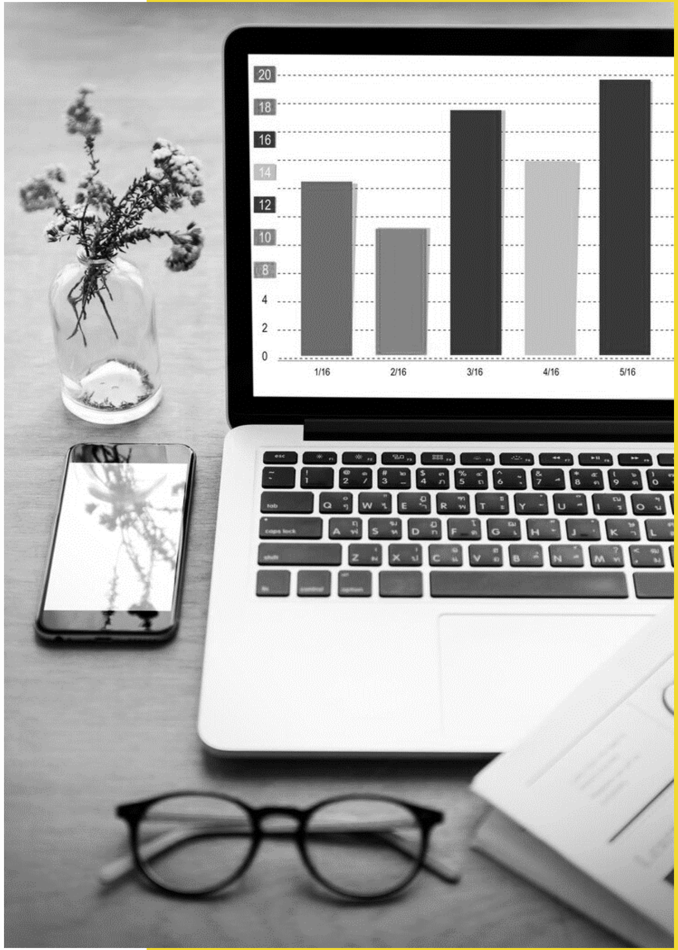
- Simple model for binary classification
- Assumes a linear relationship between the criteria and loan acceptance
- Cannot grasp complex sets, which will result in low accuracy for this task

Decision Trees

- Hierarchical structures consisting of decisions based on a specific criteria, and possible outcomes.
- Looks for best criteria for decisions
- Easy to understand
- Tailor too much to one criteria

Gradient Boosting

- Creates decision trees, where each tree corrects mistakes of the previous one
- Keeps learning by reinforcing correct decisions
- Flexible, with high prediction accuracy in complex datasets
- Difficult to tune and train



Random Forrest Classifier

Feature importance

Provides chart with top criteria that influence the outcome.

Complex, yet easy to set-up

It uses a combination of models including decision trees to capture complex non-liniear relationships, while remaining accessible and easy to maintain.

Accurate and robust to outliers

Less senstive to outliers and naturally handles categorical values without the need to actively pre-process data.

05

Model Performance

Training

Dataset that features 5,000 clients was split 70/30 while maintaining accurate Declined/Accepted ratio:

- 3,500 for training
  - Declined: 3164
  - Accepted: 336
- 1,500 for testing
  - Declined: 1356
  - Accepted: 144

Random Split was set to return same split every time to ensure accurate testing.

Personal Loan column was chosen as the criteria to check.

Other columns were selected as the criteria that can influence the state of Personal Loan (accepted vs declined).

Performance Indicators

Loan Accepted	Precision	Recall	F1-score	Instances
No	0.99	1.00	1.00	1356
Yes	0.99	0.94	0.96	144
		Totals		
Accuracy			0.99	1500
Macro Avg	0.99	0.97	0.98	1500
Weighted Avg	0.99	0.99	0.99	1500

Precision: accuracy of positive predictions.
Recall: model's ability to capture all relevant instances.
F1-score: average of precision and recall, providing a balanced metric that accounts false positives and false negatives.
Instances: number of occurrences in the dataset.
Accuracy: ratio of correctly predicted instances to the total instances.
Macro Average: average performance across different classes.
Weighted Average: average performance across different classes, considering class imbalances.

Testing

After being trained on the dataset of 3,500 instances, the model was successfully tested against 1,500 remaining instances that it hasn't seen before.

Precision: 99.33% of the time it accurately identified correct state of Personal Loan of a client (Accepted vs Declined)

Recall: 94% of the time it accurately identified clients who could accept a loan

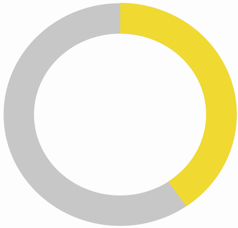
What is the difference and why?
Recall looks for ALL relevant instances, even it catches some irrelevant ones in the process.

Precision focuses on identifying how many of the relevant instances are actually correct.

The model correctly predicts whether a person accepted the loan, even if it catches some false instances 6% of the time.

Top-5 Decision Making Criteria

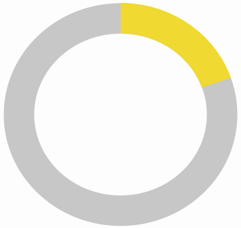
32%



Income

- Clients with income of over $100,000 are more likely to accept a loan

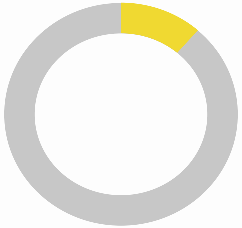
18%



Education

- Clients with post-secondary education or higher are more likely to accept a loan

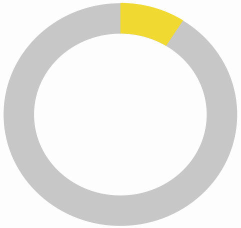
16%



CC Average

- 5.1% of income - maximum CC Average for clients who accepted a loan
- Only **ONE** client who accepted a loan **did not** have an active CC Average

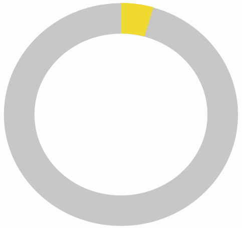
12%



Family

- Clients with 3-4 people in their immediate family are more likely to accept a loan
- *5% lower for families of 1-2*

5%

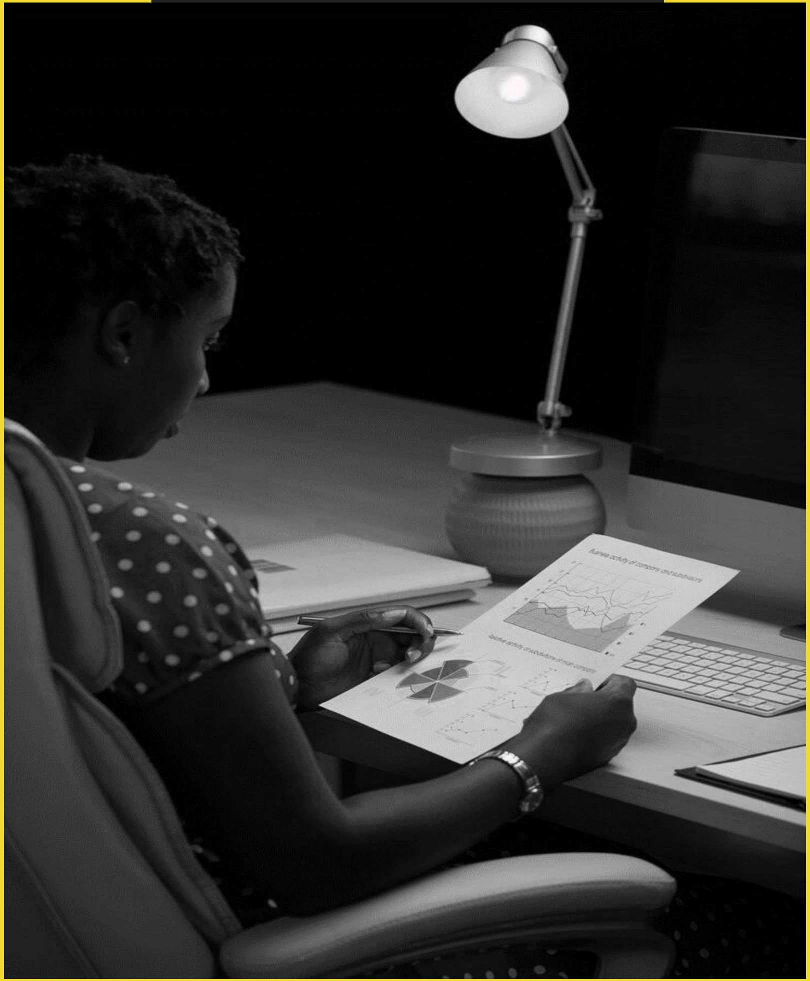


CD Account

- Clients with an active CD Account are more likely to accept a loan.
- Shown long-term commitment
- Are less likely to make rushed transfers

06

Implications



Implications on Users (Marketing Team)

- **Communication (IT):** Notify of changes in dataset ahead of time for IT to adjust the model accordingly.
- **Brand Perception:** Innovative and customer-centric bank.
- **Integration:** Understanding the information the model provides and how to integrated with the existing marketing toolkit.
- **Campaigns:** Allows to target precise group of current & potential clients. Results will need to be monitored and compared.

Implications on Clients

- **Privacy:** Personal information is being used to develop a product.
- **Target Marketing:** Tailored offers, increased communication.
- **Bias:** Clients from certain ZIP-codes may face bias.
  The model views ZIP-codes as a low-influence criteria.
- **Financial Impact:** Accepting a loan may influence their standing and credit score, especially if they are unable to pay.

Implications on Business

- **Risks:** Inaccurate predictions may result in loss of clients and money, if they are unable to repay loans. Assessing creditworthiness is the key step prior to predicting the outcome.
- **Compliance:** Ensure the model is compliant with regulations.
- **Efficiency:** Cost-effective and streamlined marketing.
- **Trust:** Strengthen the relationship with the bank. The opposite applies if the model is wrong.

Thanks!

Do you have any questions?

Reach out to **Nikita Gaidamachenko** on **Discord**