Santander Customer Transaction Prediction

Jacob Braun, Tanay Dhaka, Akshita Srivastava, Yujun Wang, Zecong Zhao

Agenda

- Problem Context
- Data Overview
- Exploratory Analysis
- Predictions & Interpretations
- Conclusion

1

Problem Context

Deriving insights from customer transactions

Business Overview

Santander is a Spanish multinational corporation bank and financial-based company operating in Europe, Asia, and North and South America

Santander continually helps its **customers**:

- understand their financial health
- discern products and services helpful for achieving their monetary goals

In this scenario, they want to identify customers who will make a **transaction** in the future, irrespective of the amount of money transacted

Analyzing customer transactions will help Santander

- Quantify a customer's willingness to transact in future
- Achieve increased customer satisfaction by curating customised products and services based on transactional behavior
- Uncover insights about customers' transactions

2

Data Overview

Deep diving into the dataset

Defining characteristics of dataset

Completely Anonymized

The dataset is completely anonymized, i.e. the features/variables have no names

Dataset with

- 200 features
- 200,000 records
- Binary variable for transaction occurrence:
  - 0: No Transaction
  - 1: Transaction

Summary

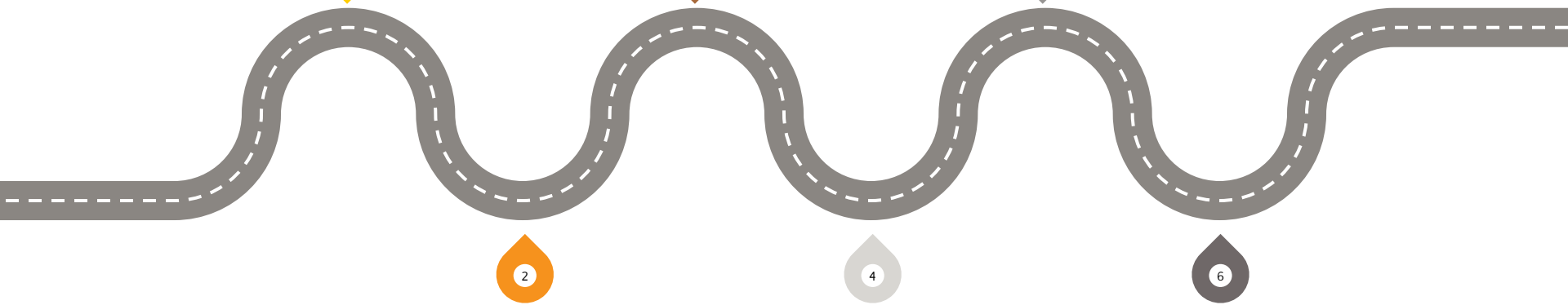
- 90% customers with no transaction
- No missing values in the dataset
- All features are numeric

3

Exploratory Analysis

A brief on data exploration, key findings and adjustments

Exploration Roadmap



Data Distribution
Almost all features were normally distributed

Duplicate values
Repetition of similar values in all columns was observed

“Magic variables”
Extra column added for every feature with real/fake flag

Correlation Analysis
No feature was correlated with another, raising suspicion about data

Identifying “Fake Data”
Data point was flagged “Real” if at least 1 feature value is unique in its own column

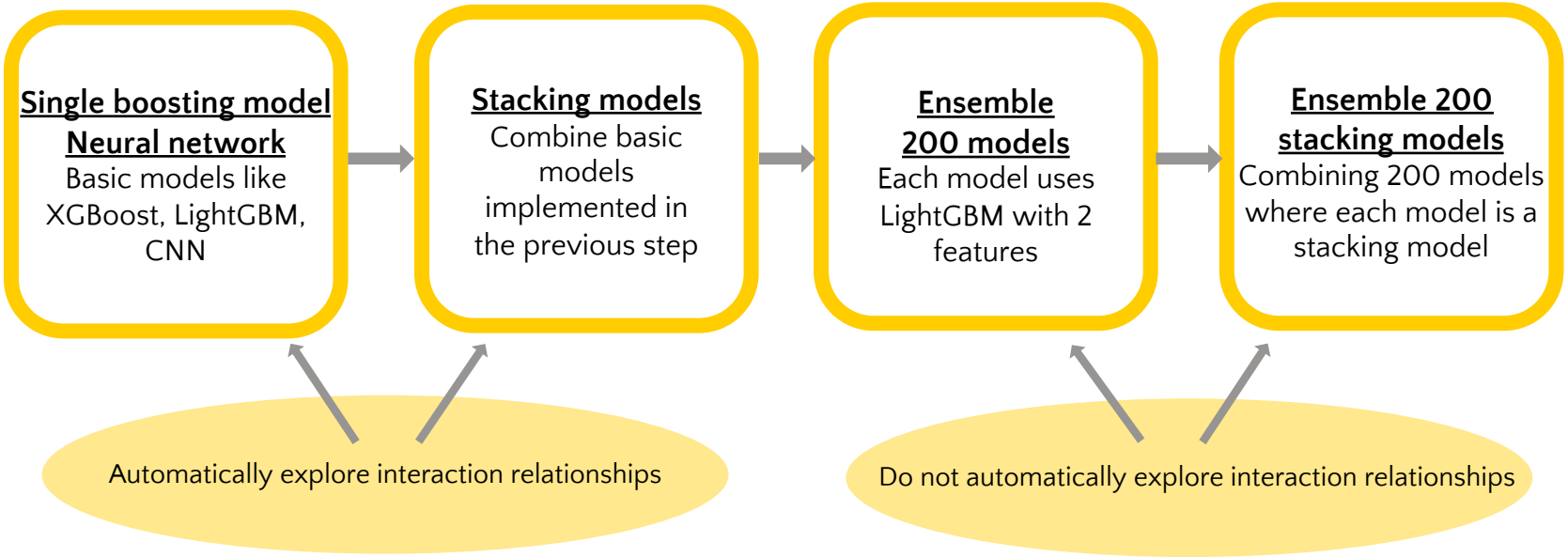
Final Analytical Data set
was built with 400 features for model training

4

Predictions & Interpretations

Various predictive models and their performances

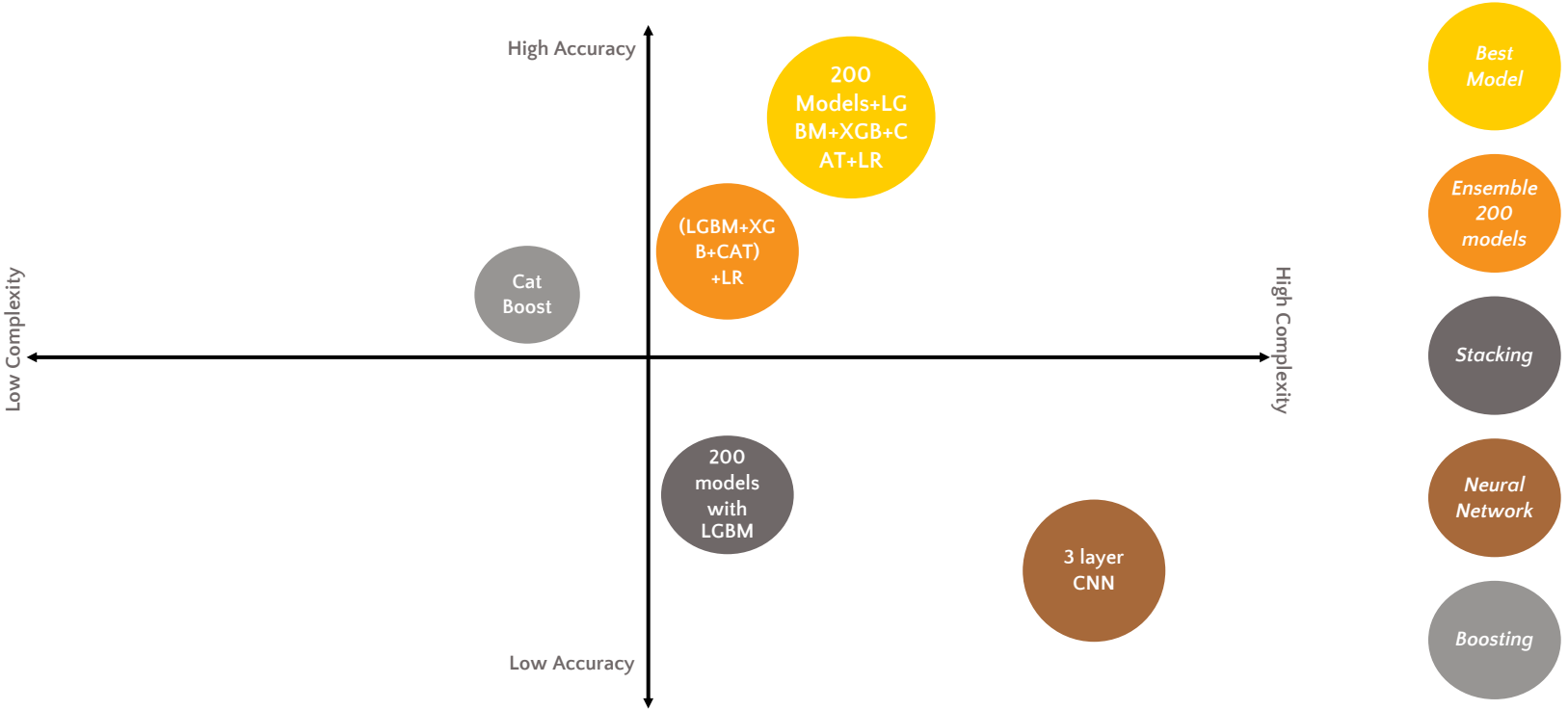
Predictive Modelling Process



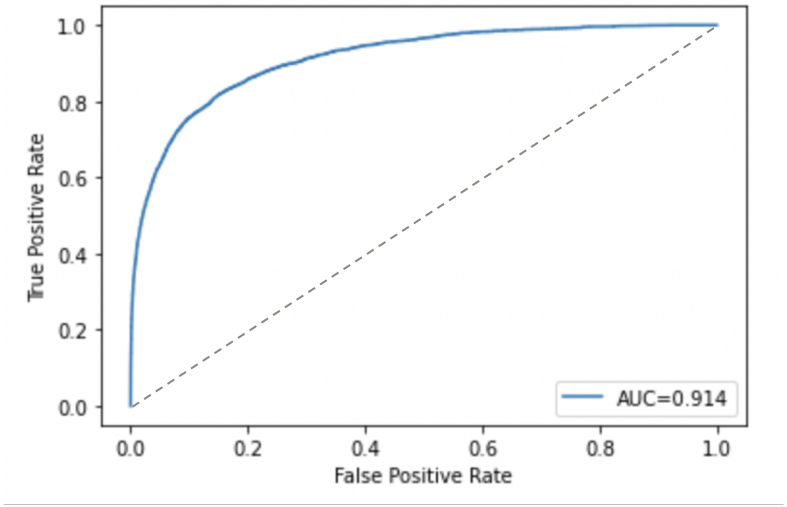
Holistic Summary of Models

Algorithm Type	*Model Type*	*Final Score*
Boosting	CatBoost	0.89484
Neural Network	CNN	0.87297
Stacking	LightGBM + XGBoost + CatBoost, meta = Logistic Regression, passthrough = true	0.90204
Ensemble 200 models	LightGBM	0.88764
Ensemble 200 stacking models	CatBoost+ LightGBM + XGBoost, meta = Logistic Regression	**0.91611**

Model Comparisons | Accuracy, Complexity & Algorithm type



Interpretations



- 91.4% of the time, our model will rank a true purchasing customer ahead of a customer that will not actually purchase

- Based on this ranking, santander can more accurately target customers who will actually have transactions and avoid wasting cost on people who will not

5 Conclusion

Summary and Next Steps

Conclusion & Next steps

- The best model can predict the possibility of a customer transacting in future with a confidence of 91.4%
- Santander can devise different strategies like:
  - increasing engagement : for customers transacting in future
  - taking retention measures : for customers not transacting in future
- Anonymized data makes it difficult to characterize the variables, better outcomes can be obtained by exploring the features

Thank You!