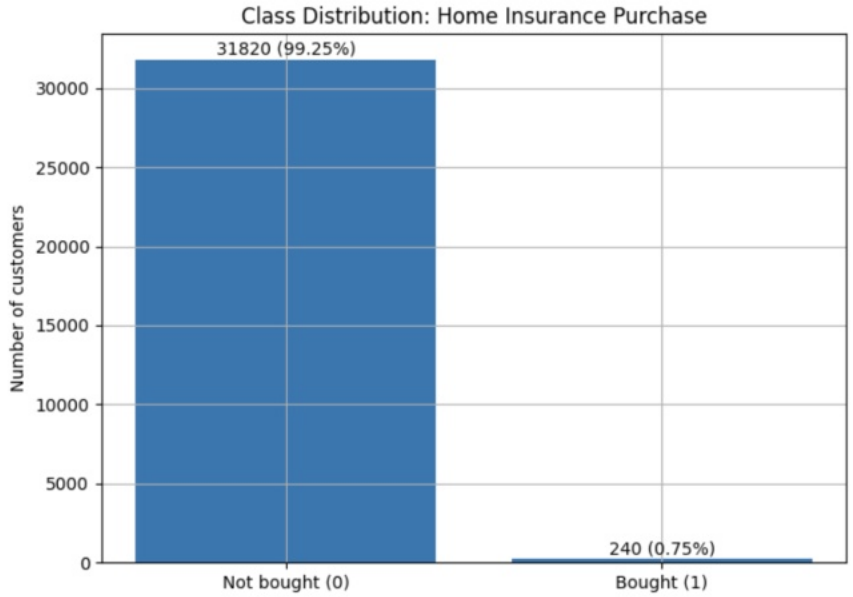
Home Insurance – Modelling Uptake from the Mortgage Book

Analyzing and predicting the uptake of home insurance among mortgage customers, using data-driven approaches and existing datasets.

Business Context

Previous marketing campaign targeted mortgage customers at random. Uptake was extremely low (≈0.75% purchase rate). We need a better selection strategy to increase conversion.



Objectives

- Low uptake in previous campaign
- Goal: rank customers effectively
- Use existing bank data only
- Ensure explainability in targeting
- Success measured by conversion rates

Data Available

- Campaign Dataset: Customer insights
- Mortgage Dataset: Employment overview
- Demographic information included
- Engagement metrics available
- Potential for data merging

Campaign Dataset Overview

The campaign dataset consists of approximately 32,000 customers, providing insights into marketing and brand variables. It serves as the foundation for understanding customer interactions and responses to the home insurance offerings.

Demographic Variables

Demographic variables include age, marital status, education level, and occupation. These factors are crucial in segmenting customers and understanding different buying behaviors related to home insurance.

Brand Attitude Metrics

Brand attitude variables such as familiarity, overall views of the bank, and interest in insurance products will help in gauging customer sentiment and potential interest in home insurance.

Outcome Variable Definition

The outcome variable is derived from the created_account field, where a 'Yes' indicates a purchase of home insurance (label = 1), while any other response is labeled as 0. This binary outcome is essential for modelling customer behavior.

Mortgage Dataset Enrichment

The mortgage dataset enhances the analysis with account-level information, including full names, dates of birth, town, PAYE status, salary bands, and employment history. This data enriches the understanding of customer profiles.

Join Strategy

To join the datasets, a strategy involving normalisation of first names and last names is employed. This ensures accuracy by trimming and lower-casing entries, followed by a left join to maintain a complete view of the campaign outcomes.

Campaign Dataset

Overview

The campaign dataset consists of approximately 32,000 customers, providing insights into marketing and brand variables. It serves as the foundation for understanding customer interactions and responses to the home insurance offerings.

Demographic Variables

Demographic variables include age, marital status, education level, and occupation. These factors are crucial in segmenting customers and understanding different buying behaviors related to home insurance.

Brand Attitude Metrics

Brand attitude variables such as familiarity, overall views of the bank, and interest in insurance products will help in gauging customer sentiment and potential interest in home insurance.

Outcome Variable Definition

The outcome variable is derived from the created_account field, where a 'Yes' indicates a purchase of home insurance (label = 1), while any other response is labeled as 0. This binary outcome is essential for modelling customer behavior.

Mortgage Dataset Enrichment

The mortgage dataset enhances the analysis with account-level information, including full names, dates of birth, town, PAYE status, salary bands, and employment history. This data enriches the understanding of customer profiles.

Join Strategy

To join the datasets, a strategy involving normalisation of first names and last names is employed. This ensures accuracy by trimming and lower-casing entries, followed by a left join to maintain a complete view of the campaign outcomes.

From Raw Fields to Modelling Features

Name Parsing

In the mortgage dataset, the full_name field is parsed into distinct components: name_title, first_name, middle_name, and last_name. This allows for cleaner joins and better handling of customer information.

Salary Engineering

The salary_band field, which is often in free-text format, is processed using a regex pipeline to derive structured salary_type categories and an estimated annual_salary_gbp. This ensures salary data is usable and standardized for analysis.

DOB & Age Calculation

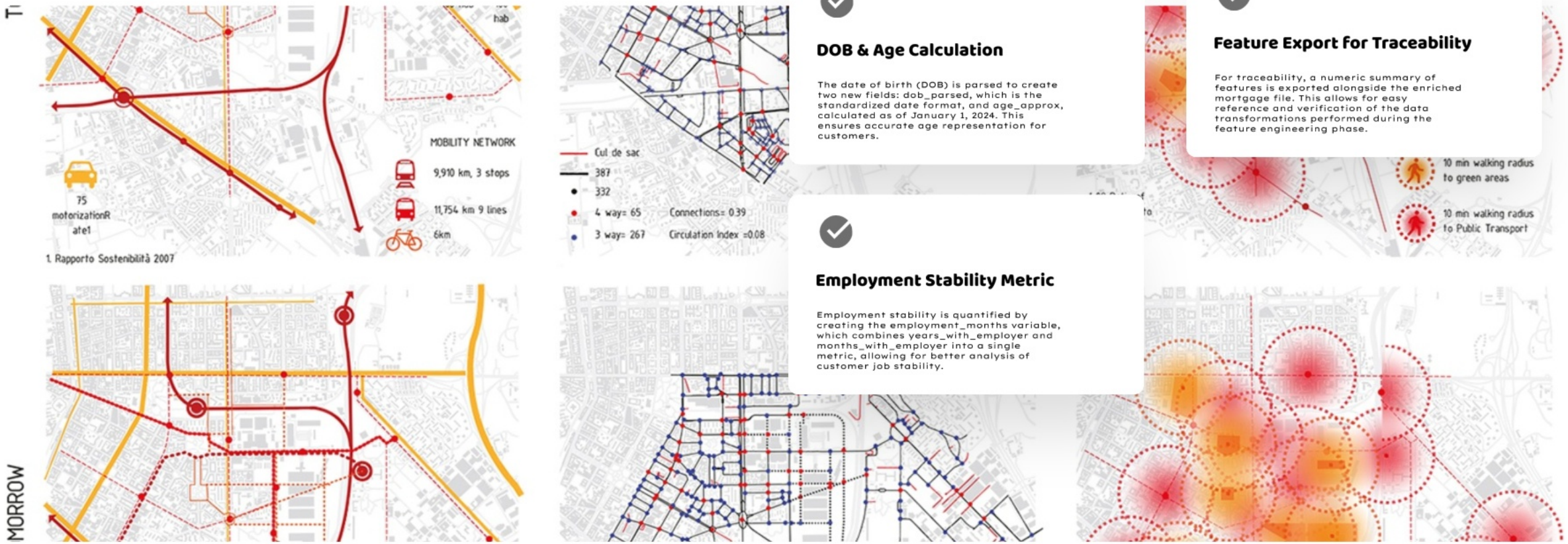
The date of birth (DOB) is parsed to create two new fields: dob_parsed, which is the standardized date format, and age_approx, calculated as of January 1, 2024. This ensures accurate age representation for customers.

Feature Export for Traceability

For traceability, a numeric summary of features is exported alongside the enriched mortgage file. This allows for easy reference and verification of the data transformations performed during the feature engineering phase.

Employment Stability Metric

Employment stability is quantified by creating the employment_months variable, which combines years_with_employer and months_with_employer into a single metric, allowing for better analysis of customer job stability.



Name Parsing

In the mortgage dataset, the full_name field is parsed into distinct components: name_title, first_name, middle_name, and last_name. This allows for cleaner joins and better handling of customer information.

DOB & Age Calculation

The date of birth (DOB) is parsed to create two new fields: dob_parsed, which is the standardized date format, and age_approx, calculated as of January 1, 2024. This ensures accurate age representation for customers.

Employment Stability Metric

Employment stability is quantified by creating the employment_months variable, which combines years_with_employer and months_with_employer into a single metric, allowing for better analysis of customer job stability.

Salary Engineering

The salary_band field, which is often in free-text format, is processed using a regex pipeline to derive structured salary_type categories and an estimated annual_salary_gbp. This ensures salary data is usable and standardized for analysis.

Feature Export for Traceability

For traceability, a numeric summary of features is exported alongside the enriched mortgage file. This allows for easy reference and verification of the data transformations performed during the feature engineering phase.

Target & class imbalance + exploratory signal

Understanding the class imbalance highlights the need for a ranking-focused approach.

0.75%

Base purchase rate

240

Total customers who bought home insurance

32,060

Total customers in the dataset

0.46%

Purchase rate for those not interested in insurance

1.14%

Purchase rate for those interested in insurance

0.75%

Base
purchase rate

240

Total customers
who bought
home insurance

32,060

Total customers in the dataset

0.46%

Purchase rate for those not interested in insurance

1.14%

Purchase rate for those interested in insurance

Target & class imbalance + exploratory signal

Understanding the class imbalance highlights the need for a ranking-focused approach.

0.75%

Base purchase rate

240

Total customers who bought home insurance

32,060

Total customers in the dataset

0.46%

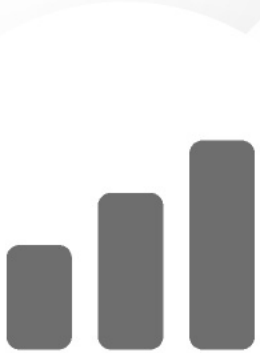
Purchase rate for those not interested in insurance

1.14%

Purchase rate for those interested in insurance

Modelling approach & evaluation

A structured approach to customer ranking and model evaluation for home insurance uptake.



Feature Set Creation

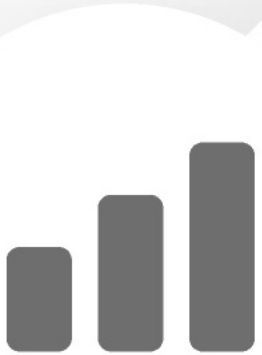
Utilize a common set of features derived from both campaign and mortgage datasets.

Attribute Management

Exclude sensitive identifiers to protect customer privacy.

Focus on Relevant Signals

Retain features related to customer behavior and financial status.



Data Splitting

Implement an 80/20 train/test split, ensuring stratification based on the target variable.

Model Selection

Evaluate three distinct models: Logistic Regression, Random Forest, and CatBoost.

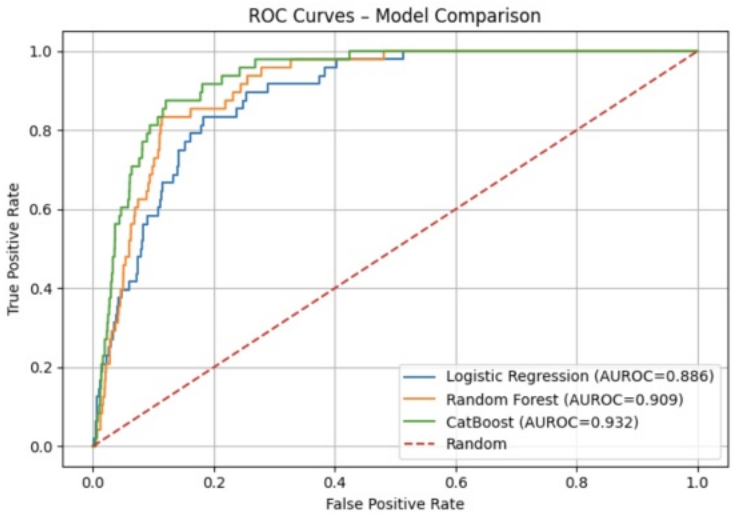
Imbalance Handling

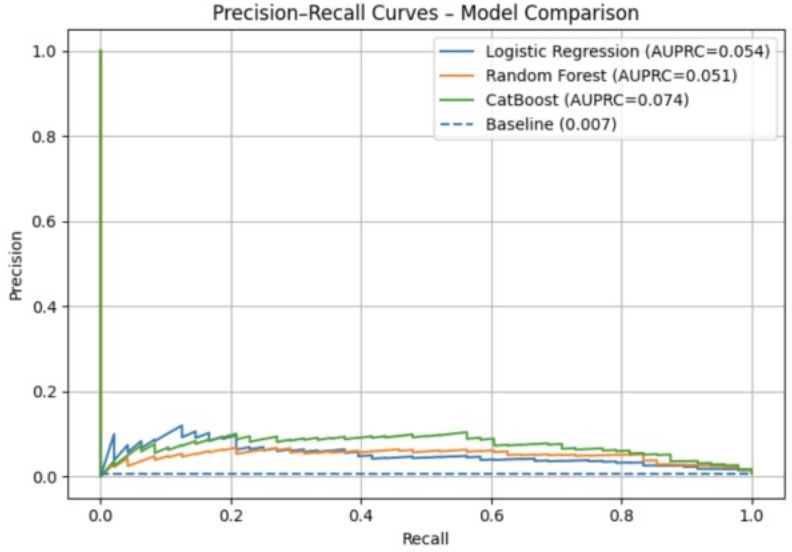
Apply class weights to all models to address class imbalance in the dataset.

Evaluation Metrics

Assess model performance using AUROC, AUPRC, Precision@K, and Lift metrics.

Model comparison





Overall Model Performance

Across the three models (Logistic Regression, Random Forest, CatBoost), overall ranking power is strong, with AUROC scores ranging from 0.886 to 0.932. This means each model reliably separates likely buyers from non-buyers despite the highly imbalanced target class (~0.7% purchase rate). CatBoost leads with the highest AUROC (0.932), showing the strongest ability to discriminate between converters and non-converters.

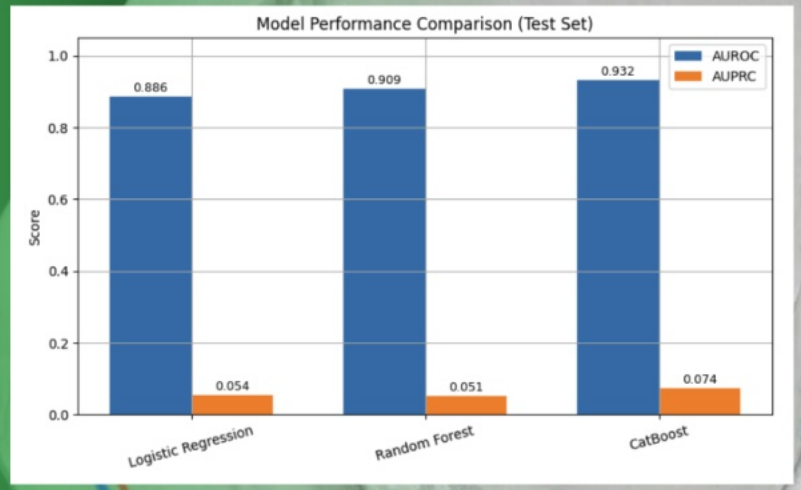
Precision and Marketing Impact

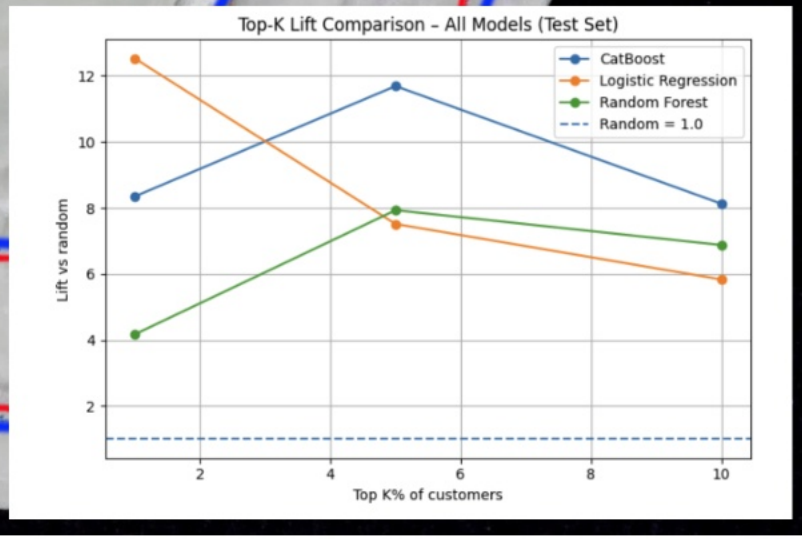
Given the extremely low base rate of conversion, Precision-Recall is the most meaningful metric for marketing. CatBoost delivers the strongest Precision-Recall performance, achieving the highest AUPRC (0.074) and exceptional Lift, reaching ~11.7x in the top 5% of customers.
This concentration of likely buyers makes CatBoost the most effective model for targeted outreach and campaign ROI.

CatBoost champion – ROC, PR & Lift

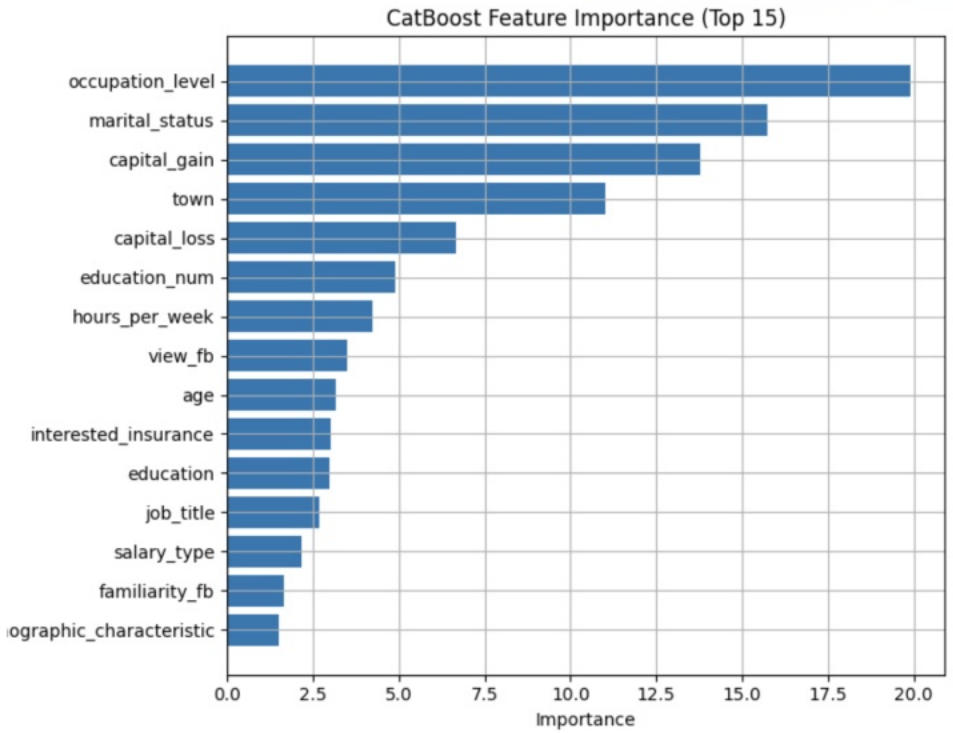
The CatBoost model demonstrates impressive performance in distinguishing between buyers and non-buyers.

We continuously monitor changes in feature distributions, missingness and demographic shifts to detect drift early.Its strong AUROC indicates excellent ranking capabilities, while the precision-recall curve shows a significant improvement over random guessing, underscoring its effectiveness in targeting potential home insurance buyers.





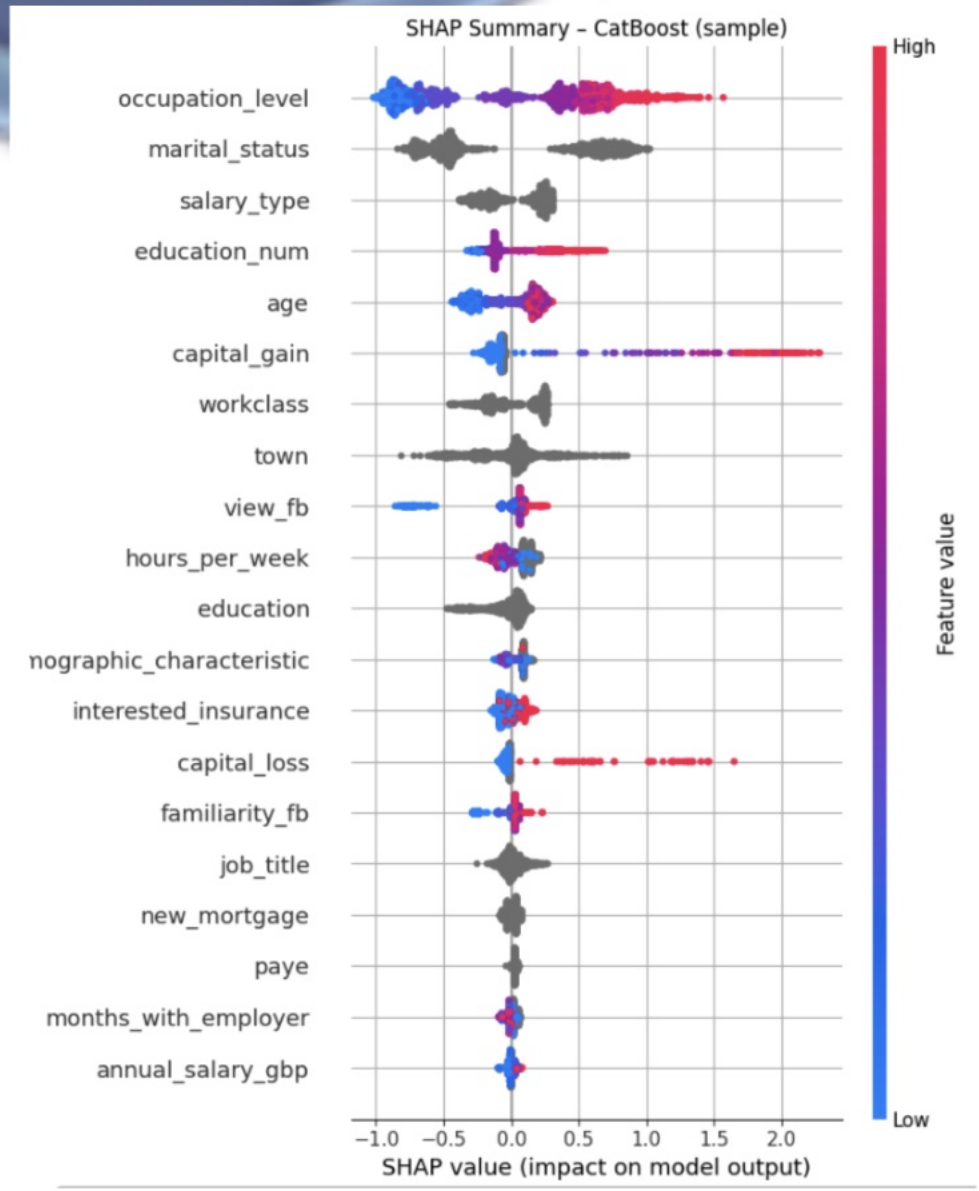
What drives uptake? (global explanations)



- Key drivers of uptake
- Occupation level matters most
- Marital status influences decisions
- Financial proxies: capital gains/losses
- Geographic patterns display trends
- Education level impacts insurance decisions
- Salary type correlates with uptake
- Brand attitudes affect purchase likelihood

Explainability in practice: reason codes & outreach lists

- Top-5 reason codes generated
- Outputs for marketing teams
- Aligns with regulatory standards
- Enables targeted customer outreach
- Supports internal governance practices



Deployment plan & monitoring

A structured approach to ensure the CatBoost model operates effectively and evolves with changing data.

Production Design

CatBoost model deployed as a batch scoring job over the mortgage book, generating ranked outreach lists regularly.

Monitoring & Retraining

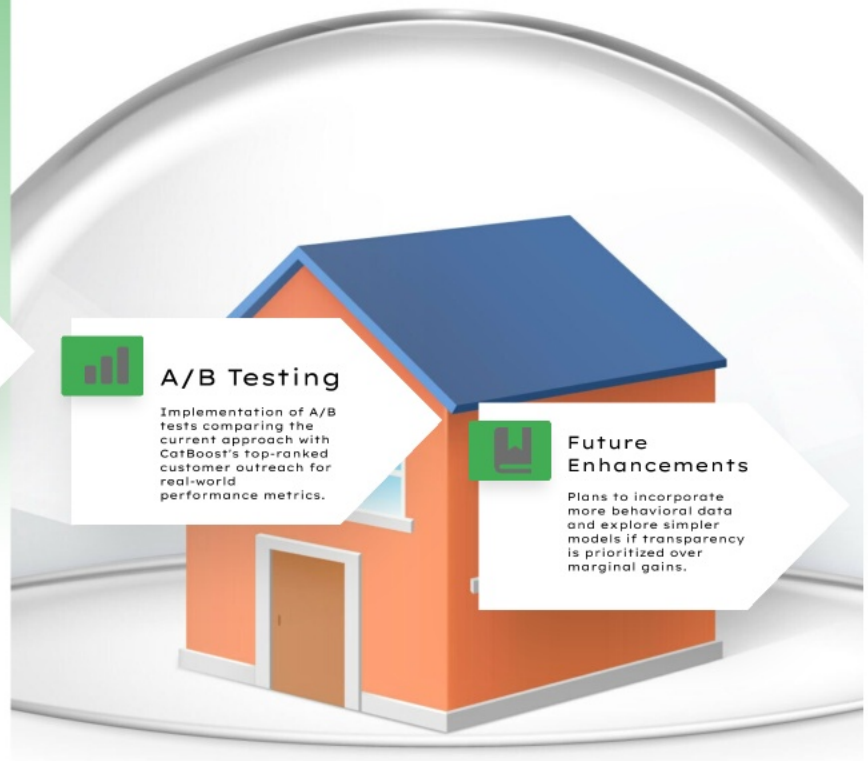
Continuous tracking of performance metrics such as AUROC, AUPRC, and Lift to ensure model accuracy and reliability.

Data & Population Drift

Regular assessment of feature distributions, missingness, and demographic patterns to identify shifts in data.

Fairness & Calibration Checks

Ongoing evaluations of model performance across diverse segments to maintain fairness and accuracy.



A/B Testing

Implementation of A/B tests comparing the current approach with CatBoost's top-ranked customer outreach for real-world performance metrics.

Future Enhancements

Plans to incorporate more behavioral data and explore simpler models if transparency is prioritized over marginal gains.

Production Design

CatBoost model deployed as a batch scoring job over the mortgage book, generating ranked outreach lists regularly.

Monitoring & Retraining

Continuous tracking of performance metrics such as AUROC, AUPRC, and Lift to ensure model accuracy and reliability.

Data & Population Drift

Regular assessment of feature distributions, missingness, and demographic patterns to identify shifts in data.

Fairness & Calibration Checks

Ongoing evaluations of model performance across diverse segments to maintain fairness and accuracy.

A/B Testing

Implementation of A/B tests comparing the current approach with CatBoost's top-ranked customer outreach for real-world performance metrics.

Future Enhancements

Plans to incorporate more behavioral data and explore simpler models if transparency is prioritized over marginal gains.

Home Insurance – Modelling Uptake from the Mortgage Book

Analyzing and predicting the uptake of home insurance among mortgage customers, using data-driven approaches and existing datasets.

Take this with you.
Revisit anytime.

Missed something? Want to explore further?
Scan or click below to open this presentation.
Anytime, anywhere.

View presentation