THE HARTFORD

Final PROJECT presentation

Hartford GROUP 3

SALISA ALMEIDA

PETER HOFSTEDT

JATIN BOMRASIPETA

PRESENTATION AGENDA

1

Project Overview

Business Problems

Key Insights

2

Data Overview

Challenges on Modeling

3

Two Step Modeling Approach

Step 1

Step 2

4

Customer Segmentation and Retention Strategies

Next Step

PROJECT OVERVIEW

The primary objective

- Analyze factors influencing customer retention

- Develop predictive models that can help identify at-risk customers

- Set strategies ideas to improve retention rates.

BUSINESS PROBLEM

Can the likelihood of a customer canceling their auto insurance be estimated?

Additionally, can we determine the specific type of cancellation they will choose, such as midterm or flat cancellation?

KEY INSIGHTS

MODEL

Performance: LightGBM showed the highest accuracy on Test set.
Precision = 0.84 for renewals
0.82 for cancellations
AUC = 0.70

Feature Engineering : Created 5 new features for step 1 model that and reduce dimensionality for improved model performance.

AS BUSINESS

Customer Segmentation: Identified four distinct customer segments with varying loyalty levels and retention needs.

Targeted Retention Strategies: Developed tailored strategies for each segment to improve retention rates and customer satisfaction.

THE HARTFORD

DATA OVERVIEW AND MODEL CHALLENGES

1. Data Cleaning and Enhancement
- Before Cleaning: 264,963 rows by 54 Columns (strategy) (based needs)
- After Cleaning: 264,963 rows by 57 Columns
  - Removed 2 variables, created 5 variables

2.Preparing for Modeling
- Label Encoding (categorical-numerical)
- Checking correlation between best predictors variables
- Checking correlation with STATUS

- Keep a good Varience = imputation, transformed categ to num and creation of variable

3.imbalanced Classes:
- STATUS:
  - 0 = Not Cancelling - 84.6%
  - 1 = Mid-Term Cancellation - 8.2%
  - 2 = End-Term Cancellation - 7.3%

To try model Imb enviroment

4.Balancing Techniques
- SMOTE
- Downsampling

TWO STEP MODELING APPROACH

Divide the target into two steps

Model in 2 ways

STEP 1: Predict customer cancellation (Class 0 or Class 1 and 2)

STEP 2: Predict mid-term and end-of-term cancellations probabilities (Class 1 or Class 2)

Metrics: We focused on precision because it helps accurately identify real cancellations and minimize false positives in predicting customer cancellations. ROC-AUC was used too as the data imbalance was severe.

THE HARTFORD

CHOOSING OUR FEATURE SELECTION METHOD FOR STEP 1

- Recursive Feature Elimination
- Principal Component Analysis
- Feature Importance ranking
- No classifier Feature selection
  - Correlation with STATUS thresholds and correlations with other variables

The best method for feature selection that we found was **Feature Importance Ranking**

FINDING THE BEST MODEL TYPE: STEP 1

We tested different models on the full dataset to choose the best model

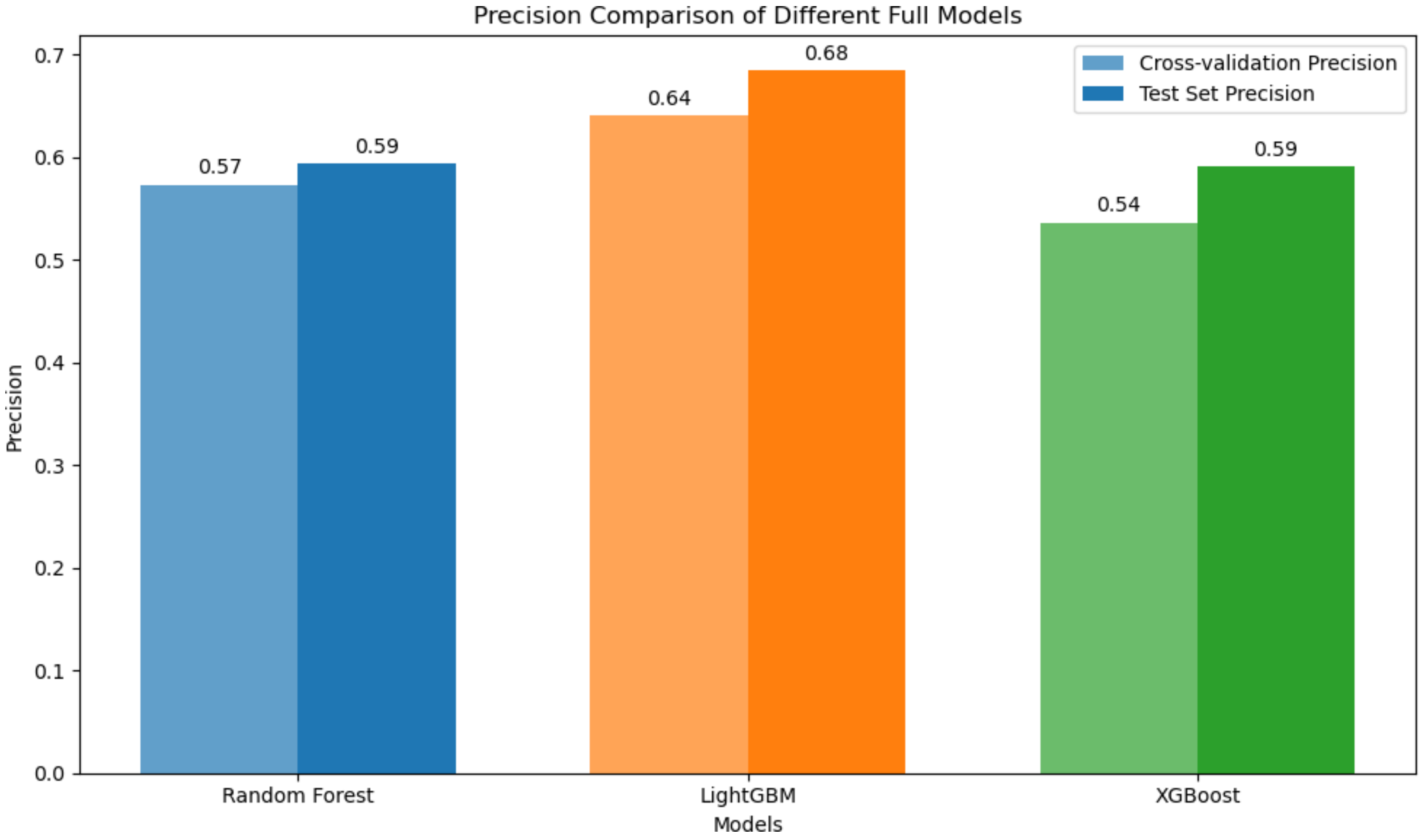
- Random Forest
- XGBoost
- LightGBM

We found that **LightGBM** had overall higher accuracy and better predictive power than the other models

LightGBM
had 68% precision for cancellations on the full model

THE HARTFORD

COMPARISON OF DIFFERENT STEP 1 MODELS ON THE FULL DATASET

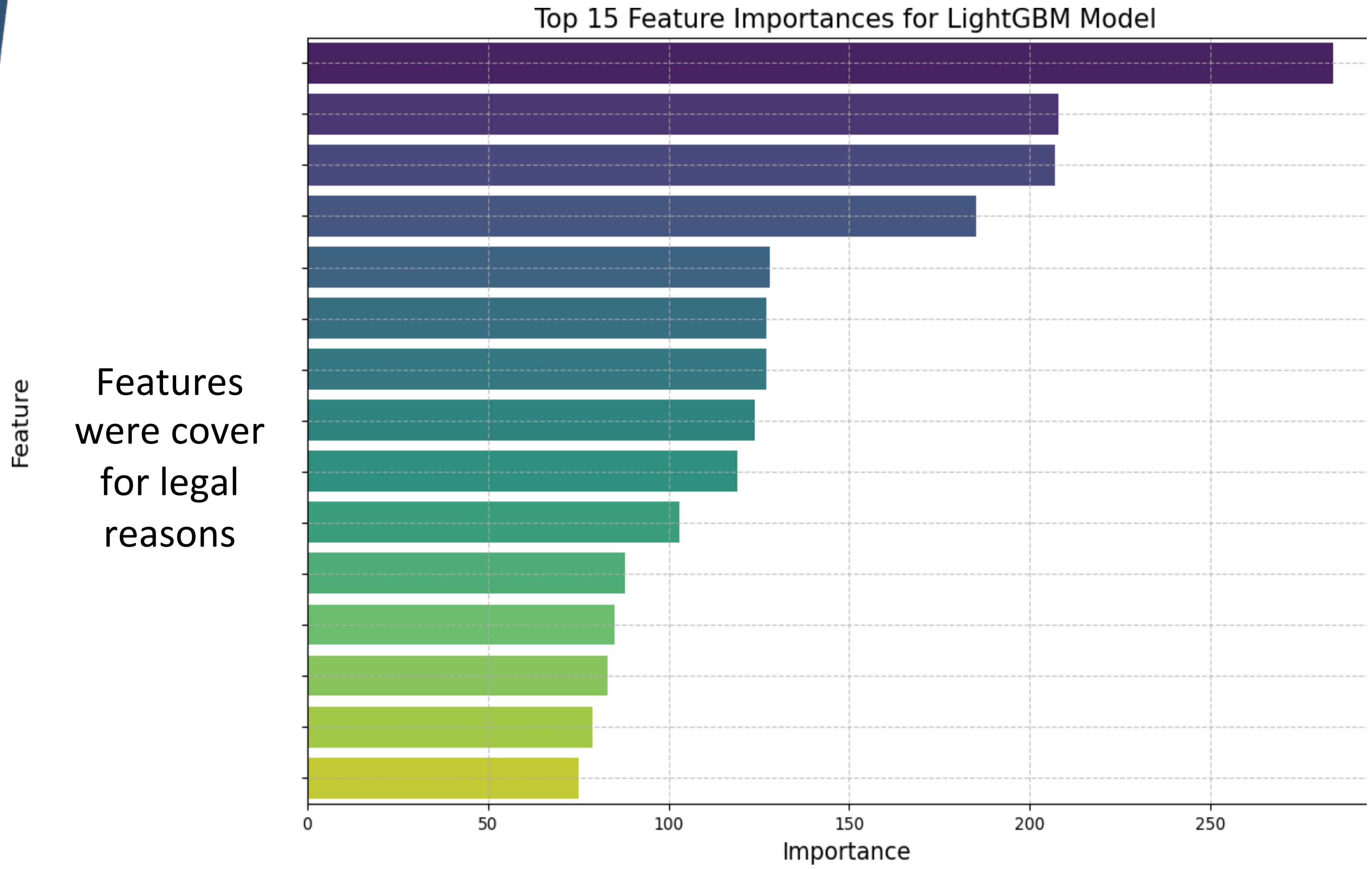


LightGBM had a better performance using all variables.

CV = 0.64
Test = 0.68

THE HARTFORD

FEATURE IMPORTANCES



Features were cover for legal reasons

LightGBM feature selection technique was used for selecting the best 15

THE HARTFORD

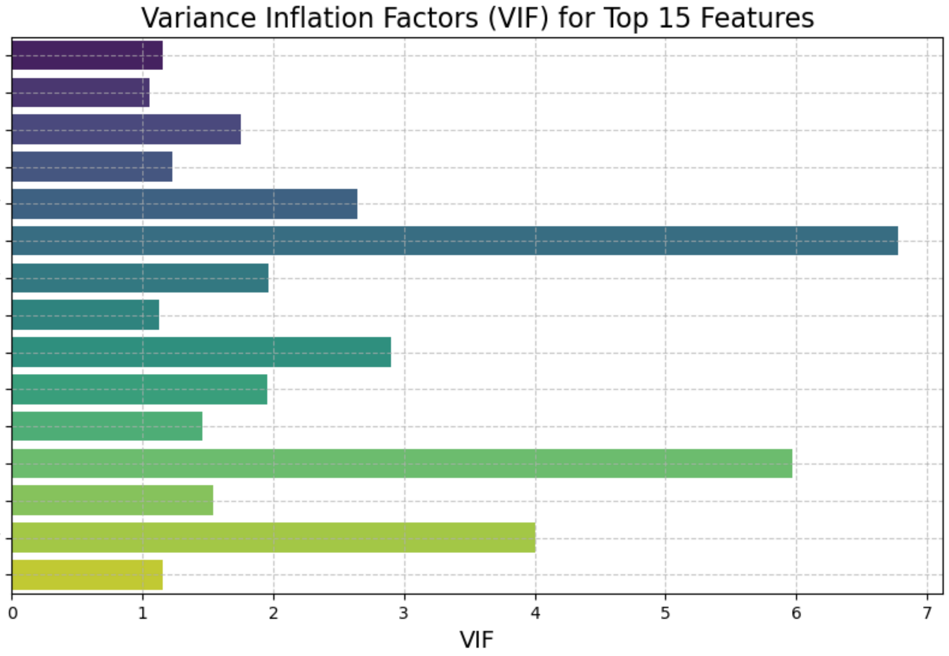
COREELATION AND VIF

15 Best Variables Cleaning

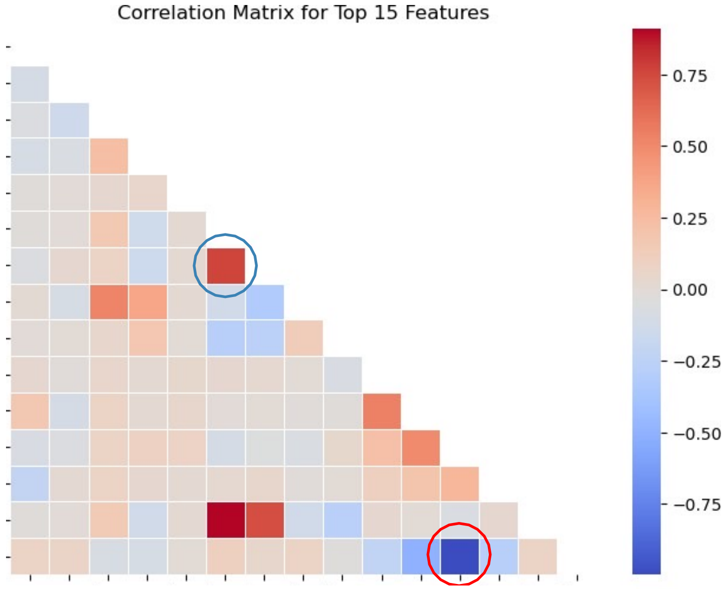
- One was removed for correlation.
- 2 Removed for VIF and high correlation.

Feature

Features were cover for legal reasons



Features were cover for legal reasons



Features were cover for legal reasons

DATA ENHANCEMENT

- Create new variables that show important business statistics.
- Increased robustness
- Reduce the complexity of models and improve efficiency
- Ensure model accuracy

FEATURE ENGINEERING FOR ENHANCED MODEL PERFORMANCE

NEW FEATURES

PREMIUM_CHANGE: Absolute change in renewal premium over two terms.

DRVR_AGE_DIFF: Difference between the oldest and youngest drivers on the policy.

BILL_PREM_DIFF: Difference between the total annual bill premium amount and the average term premium.

AVG_COMP_BI_RTNG_SYM: Average of comprehensive and bodily injury rating symbols.

VEHICLE_AGE: Age of the vehicle calculated from the model year.

Variables used to create new features were combined to form new ones to avoid multicollinearity and higher prediction power.

STEP 1 FINAL MODEL FEATURE SELECTION

Original Features

New Features

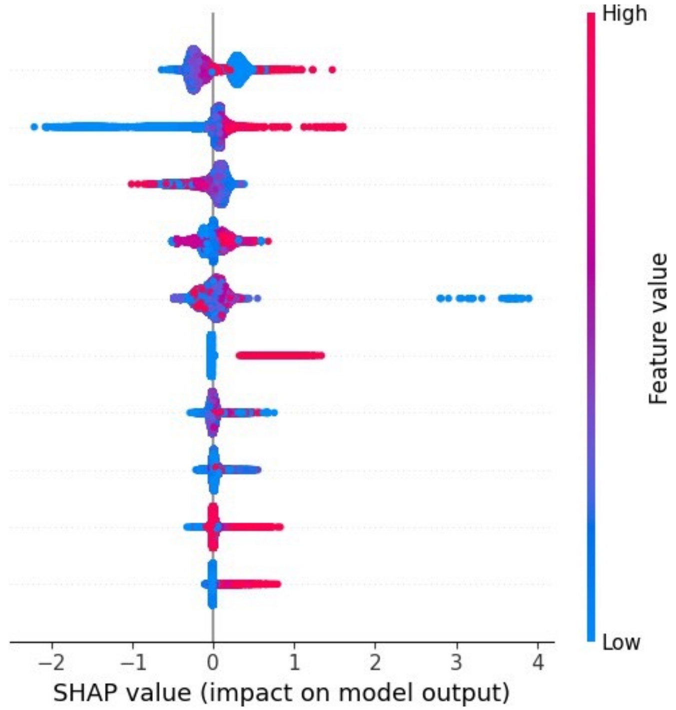
Features were cover for legal reasons

CHECKING SHAP VALUES FOR THE FIRST MODEL RESULTS

Positive SHAP values indicate an increase in the probability of cancellation and vice-versa.

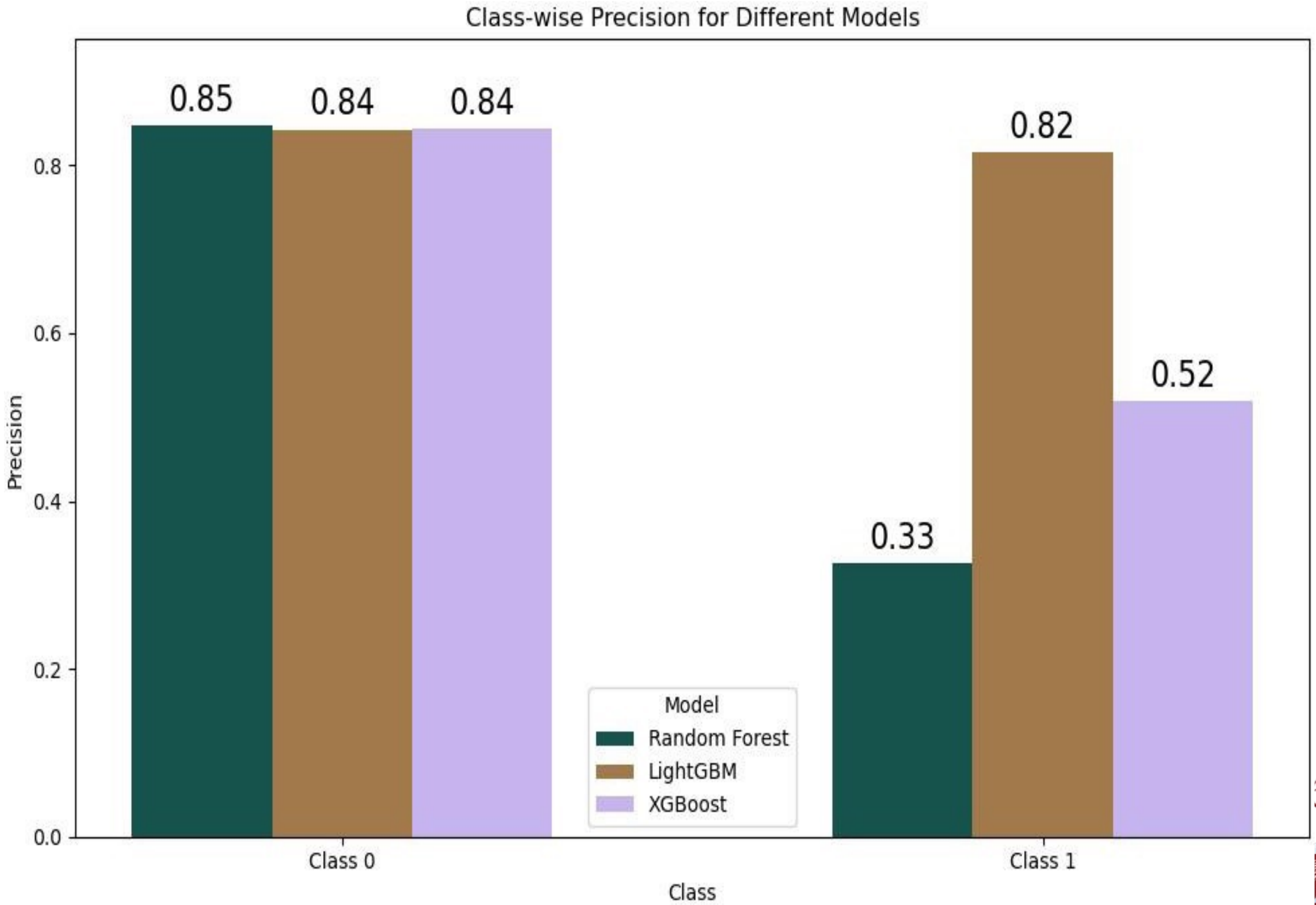
SHAP Values best features for predicting cancellations: BILL_PREM_DIFF, VEHICLE_AGE and YEARS_WITH_HIG

Features were cover for legal reasons



STEP 1 FINAL MODEL PERFORMANCE

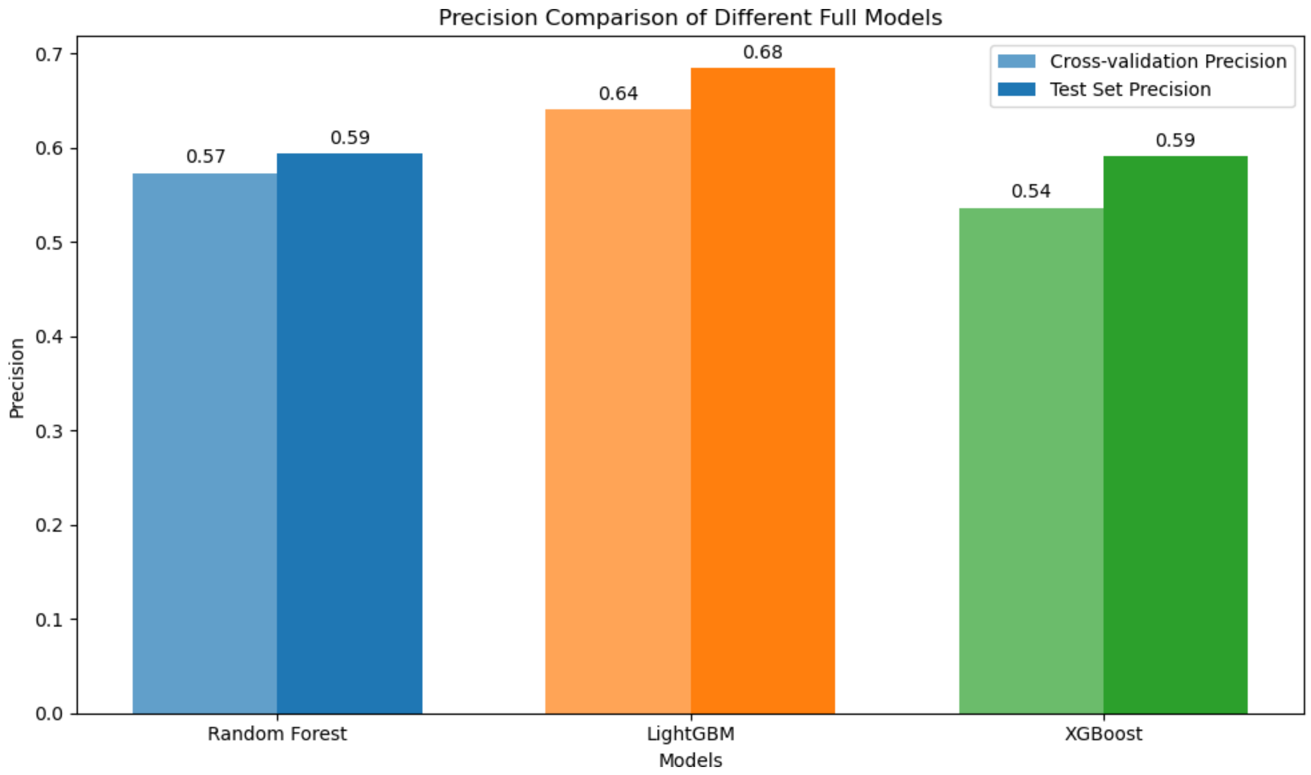
Step 1 Final Model Performance comparison of cancellations between models.
Precision class 0 = 84, 1 = 82



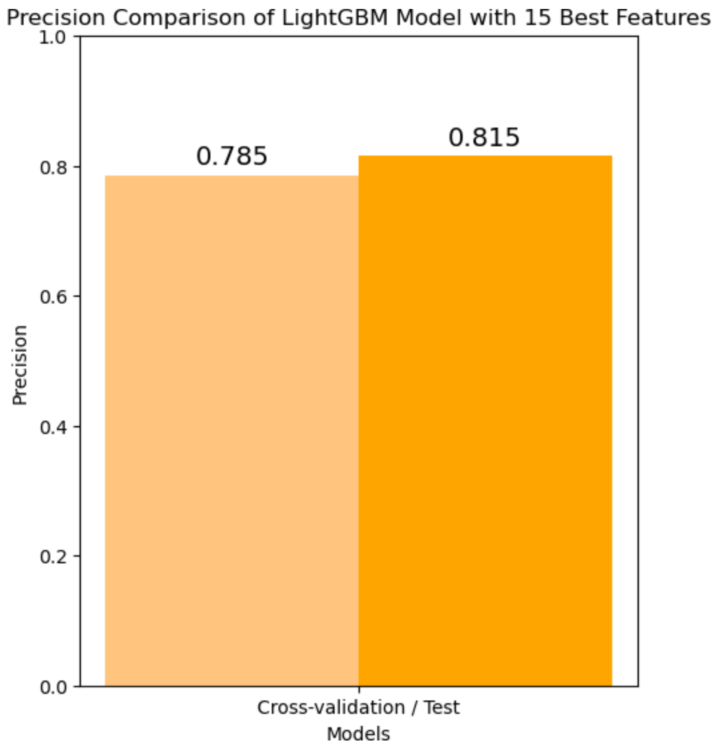
THE HARTFORD

STEP 1 MODEL RESULTS ON THE TEST SET

Full Dataset Prediction



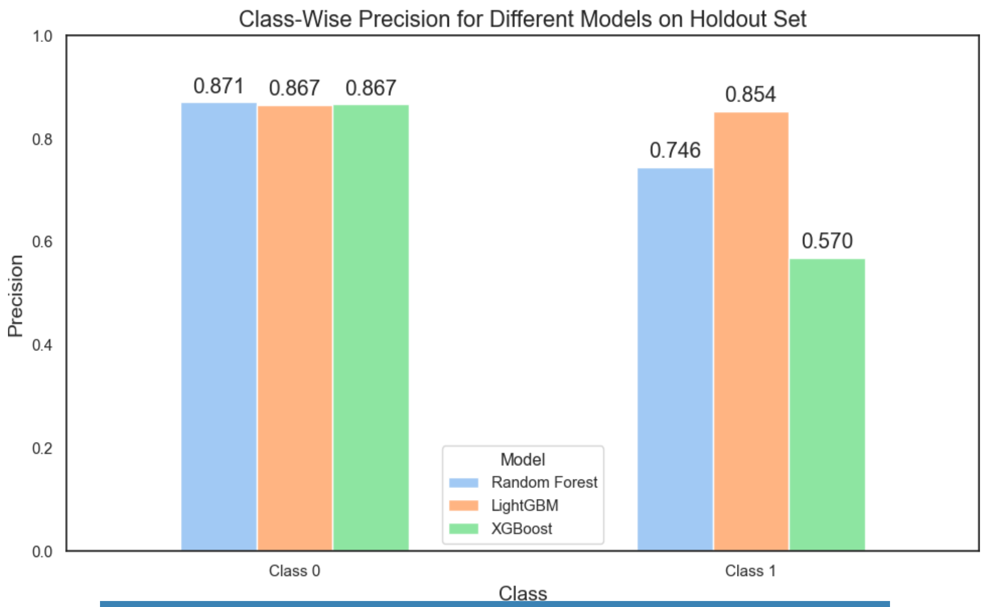
Step 1 Final Model Prediction



LightGBM had a better performance using all variables and reached 0.82 on Test set using the 10 best variables for prediction.

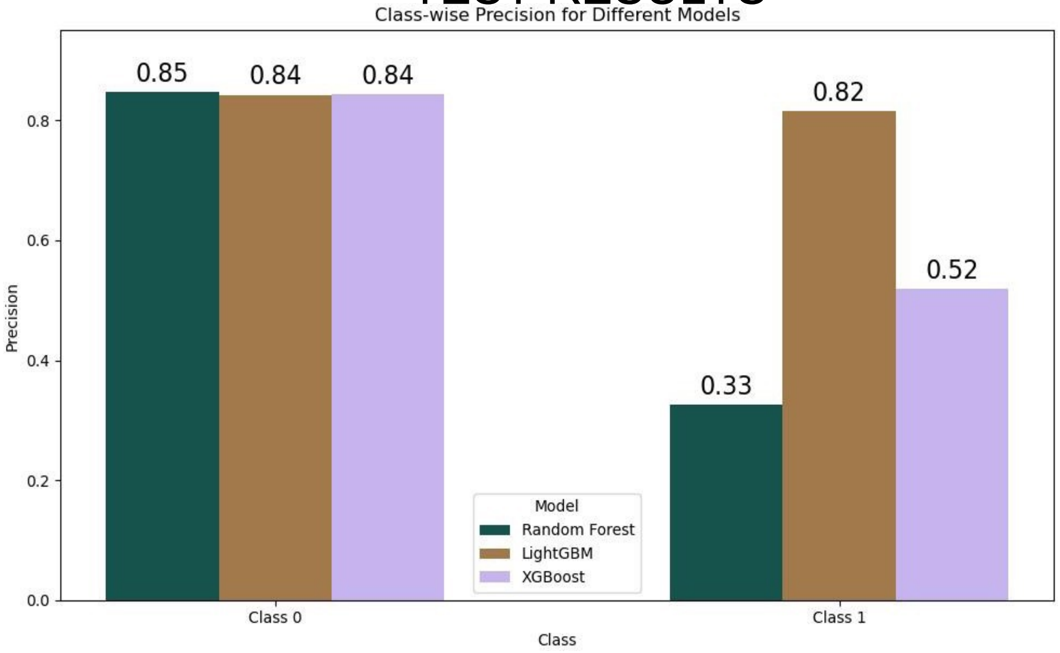
STEP 1 MODEL - HOLDOUT SET VS TEST SET

HOLDOUT RESULTS



The consistent performance across the test set and holdout set indicates that the model is not overfitting

TEST RESULTS



The model showed a slightly better performance on the holdout set.

STEP 1: 0THER APPROACHES

Techniques we tried to implement to improve our prediction results

- Weighting
  - Increases other metrics at the cost of precision

- SMOTE
  - No changes on classes precision

- Lasso and Ridge regularization
  Did not change precision

- PCA
  Class 1 was overestimated

- Learning Rate Scheduler

STEP 2 MODEL PREDICTION

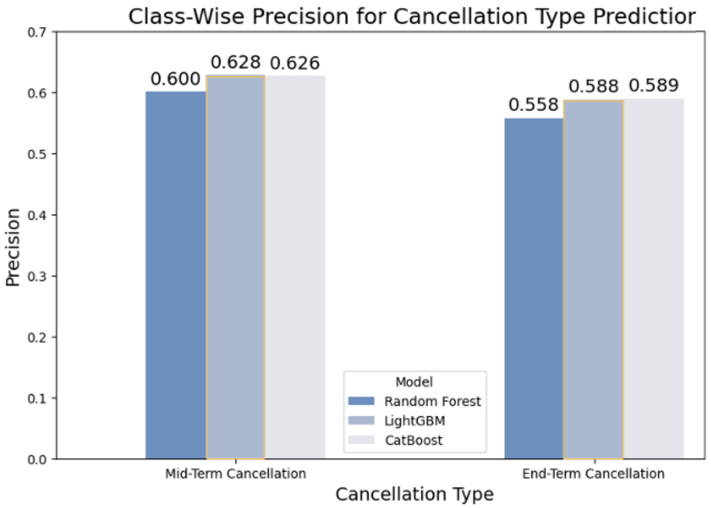
Progress and Challenges

Target Variable Creation: Filtered out non-cancellations and used STATUS column for cancellation type.

Feature Selection: Used LightGBM, identified ZIP Code and Policy Effective Date as top features. Various feature selection techniques were applied, but none significantly improved model performance.

Challenges Faced: Insufficient data for minority classes, impacting performance despite various techniques.

STEP 2 Model Results



Precision didn't meet levels we found acceptable:
Class 1: 63% Class 2: 59%

STEP 2 Cancellation type model: Best results were achieved by 15 top feature importance selection using LightGBM classifier

CUSTOMER SEGMENTATION OVERVIEW

We focused on **3 KEY FEATURES** from the **STEP 1 MODEL** to divide the customers into **4 SEGMENTS**

Features were cover for legal reasons

METHOD USED: KMEANS CLUSTERING

	Premium Difference	Vehicle Age	Avg Years with insurance
SEGMENT 0	Moderate	Older Vehicles	7+
SEGMENT 1	Very Low	Oldest Vehicles	6+
SEGMENT 2	Highest	Newest Vehicles	9+
SEGMENT 3	Relatively High	Relatively New	8+

THE HARTFORD

SEGMENT CHARACTERISTICS AND STRATEGIES

	SIZE	CUSTOMER LOYALTY	STRATEGIES
SEGMENT 0	100,000+	HIGH (LONG TERM CUTOMERS)	LOYALTY PROGRAMS & CUSTOMER FEEDBACK
SEGMENT 1	100,000+	MODERATE (POTENTIAL FOR GROWTH)	ENGAGEMENT CAMPAIGNS & INCETIVES ON RENEWALS
SEGMENT 3	ABOUT 5,800	VERY HIGH	PREMIUM SERVICES & PERSONALIZED OFFERS
SEGMENT 4	AROUND 40,000	HIGH BUT NEEDS ENGAGEMENT	REGULAR CHECK INS & TARGETED DISCOUNTS

NEXT STEPS (if we had more time)

1.**Gather More Data**: Collect additional information on customers who cancel mid-term to better understand their characteristics and reasons for cancellation.

2.**Integrate External Data Sources**: Incorporate external data sources, such as economic indicators and competitor actions, to enhance the model’s predictive power and provide a more comprehensive view of the factors influencing customer cancellations.

3.**Continue the Development of Step 2 Model**: Find and implement new techniques that could help increase the predictive power of the step 2 model.

4. **Adding both model** : Create a new model adding step one and 2

Thank you

Q\A