Predicting Auto Insurance Outcomes with Machine Learning

Agile Insurance Consultancy

Yuma Teshirogi, Joey Shum, Noah Nguyen, Quang Bui

01.

Background and objectives

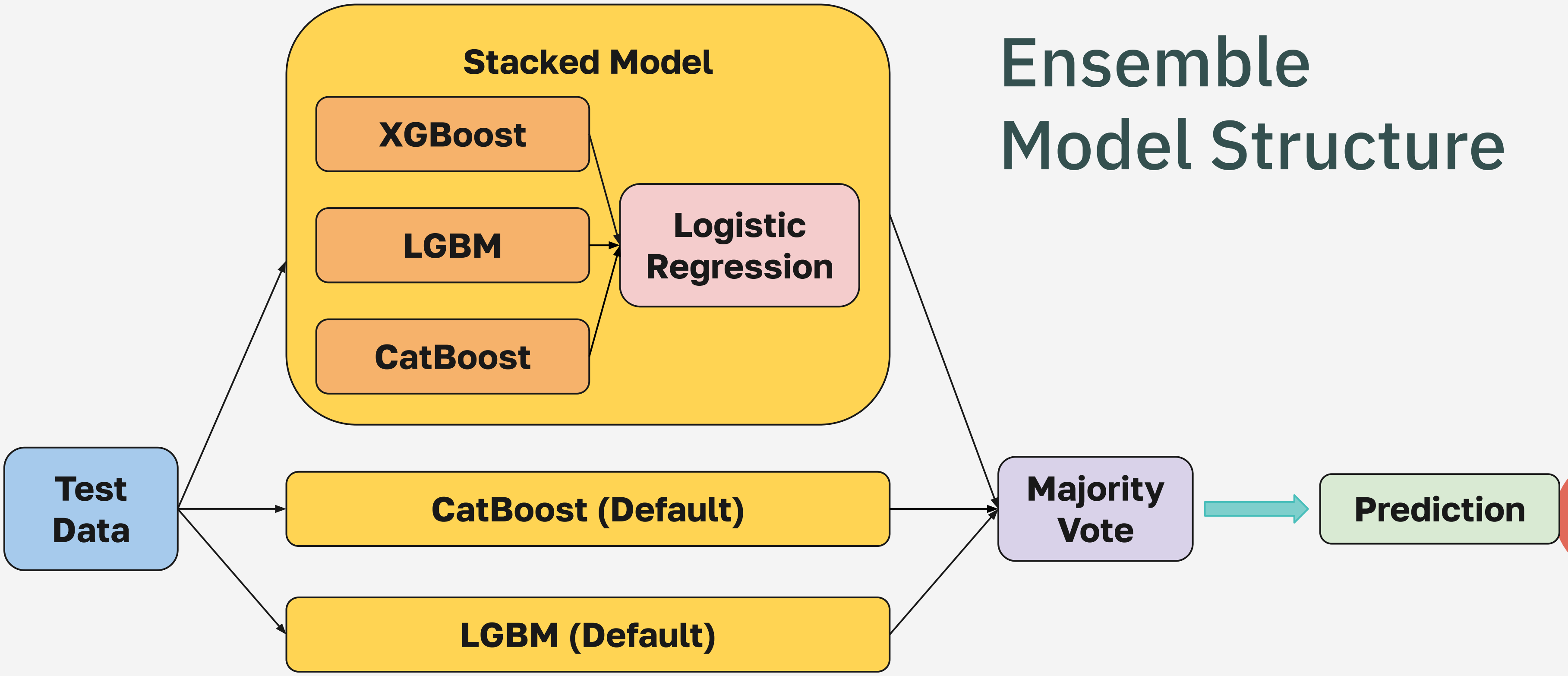
Our objective

- Objective: Predict auto insurance claims over $1000
- Model construction and analysis based on 7,290 records of auto insurance claims with 29 total features
- Focused on developing a model with both high accuracy and strong explainability

02.

Technical analysis

Ensemble Model Structure



Feature Engineering

Derive Average Metrics

- *Avg CLV = CLV / Months since Policy Inception*
- *Avg Complaints = Complaints / Policy*

Reformat Policy Age

- Given date format MM/DD/YY
- Converted to a numerical column for easier interpretation

Create Expected Claim Size

- Created Expected Claim Size (ECS) utilizing existing features (CLV, Months since Last Claim, Months since Inception, Number of Policies)
- *ECS = CLV * (MSLC / MSI) * log(1 + Policies)*

Data Processing

- Index Column
- One Hot Encoding
- Standard Scaler
- Robust Scaler

Feature Selection

- SHAP Chart
- Feature Importance Chart
- Exploratory Data Analysis (EDA)

Hyperparameter Tuning

- learning_rate
- max_depth
- n_estimators
- Reg_alpha (l1)
- Reg_lambda (l2)

Data Processing

- Index Column
- One Hot Encoding
- Standard Scaler
- Robust Scaler

Feature Selection

- SHAP Chart
- Feature Importance Chart
- Exploratory Data Analysis (EDA)

Hyperparameter Tuning

- learning_rate
- max_depth
- n_estimators
- Reg_alpha (l1)
- Reg_lambda (l2)

Algorithm \ Metric	Recall	F1 Score	ROC-AUC	Accuracy
CatBoost	0.567	0.662	0.775	0.934
XGBoost	0.559	0.642	0.760	0.929
LightGBM	0.565	0.647	0.773	0.930

- Metrics and learning curves are obtained using 5-folds stratified cross-validation
  - Recall, F1 score, ROC-AUC, Accuracy
- **Learning Curves**: training size progressively increases from 10% to 100%
- **Confusion Matrix**: based on result of first fold

XGBoost

		Predict	
		0	1
Actual	0	1262	30
	1	77	89

CatBoost

		Predict	
		0	1
Actual	0	1257	35
	1	77	89

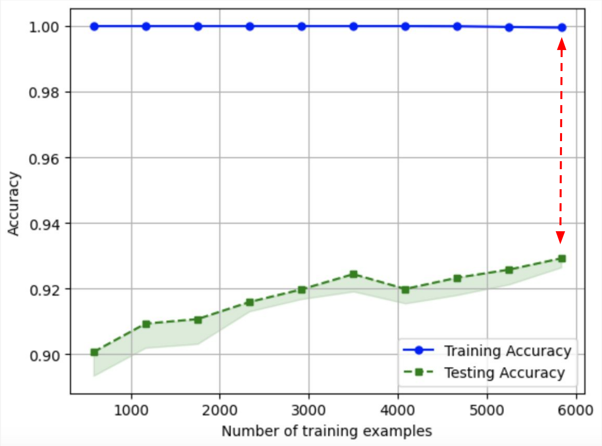
LightGBM

		Predict	
		0	1
Actual	0	1269	23
	1	76	90

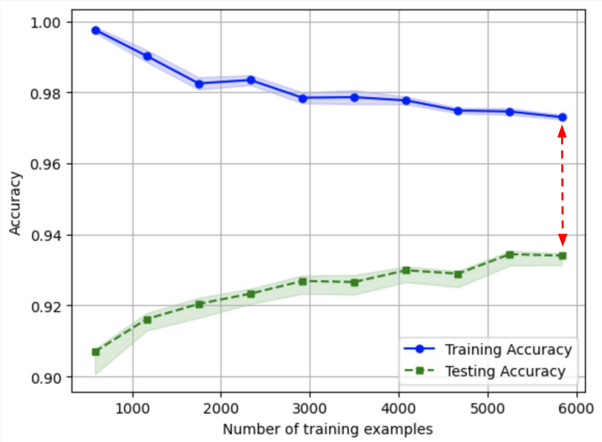
Algorithm \ Metric	**Recall**	**F1 Score**	**ROC-AUC**	**Accuracy**
CatBoost	0.567	0.662	0.775	0.934
XGBoost	0.559	0.642	0.760	0.929
LightGBM	0.565	0.647	0.773	0.930

- Metrics and learning curves are obtained using 5-folds stratified cross-validation
  - Recall, F1 score, ROC-AUC, Accuracy
- **Learning Curves**: training size progressively increases from 10% to 100%
- **Confusion Matrix**: based on result of first fold

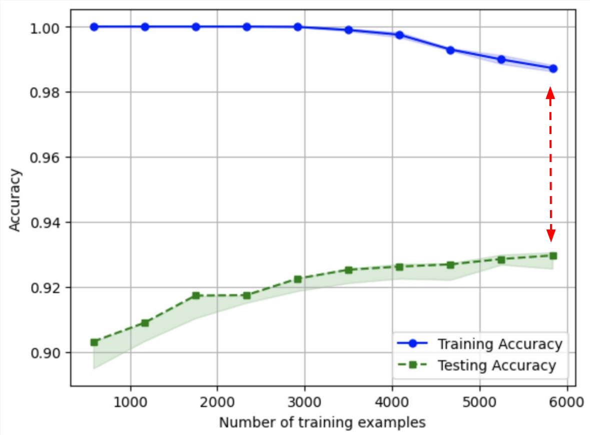
XGBoost



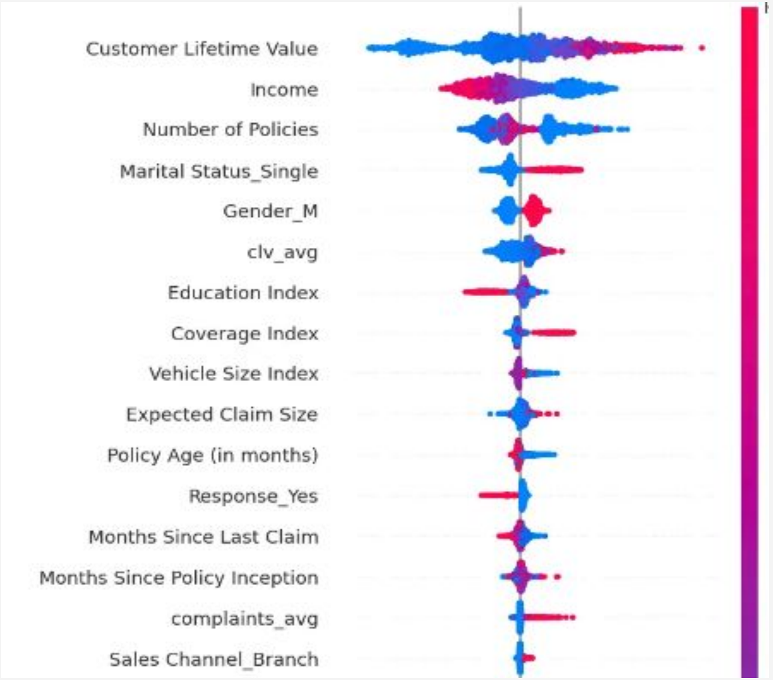
CatBoost



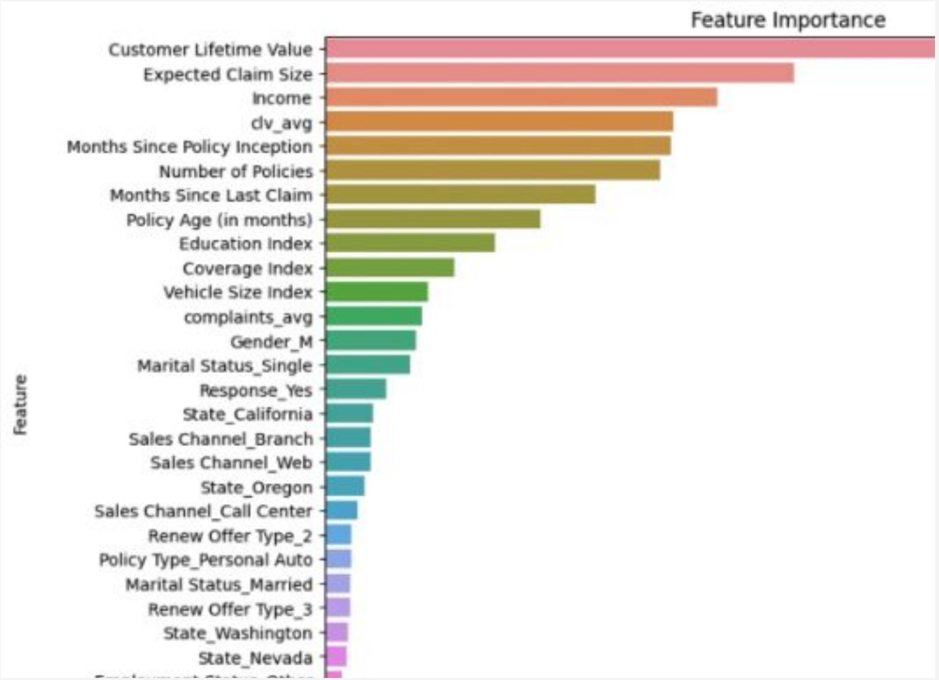
LightGBM



SHAP chart



Feature importance chart



Irrelevant features and features with low importance are removed to minimize noise.

Data Processing

- Index Column
- One Hot Encoding
- Standard Scaler
- Robust Scaler

Feature Selection

- SHAP Chart
- Feature Importance Chart
- Exploratory Data Analysis (EDA)

Hyperparameter Tuning

- learning_rate
- max_depth
- n_estimators
- Reg_alpha (l1)
- Reg_lambda (l2)

Data Processing

- Index Column
- One Hot Encoding
- Standard Scaler
- Robust Scaler

Feature Selection

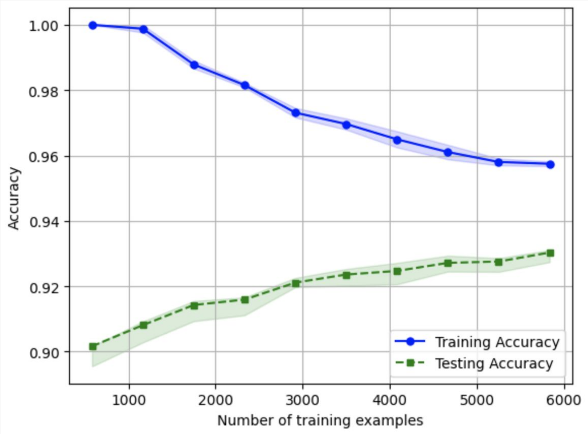
- SHAP Chart
- Feature Importance Chart
- Exploratory Data Analysis (EDA)

Hyperparameter Tuning

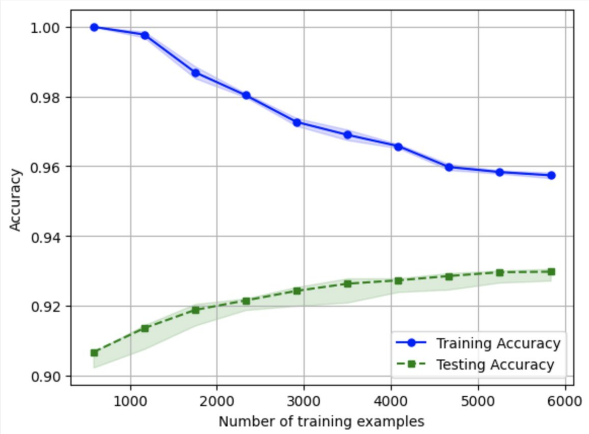
- learning_rate
- max_depth
- n_estimators
- Reg_alpha (l1)
- Reg_lambda (l2)

After tuning, we achieve best F1 scores of **0.6435** for XGBoost, **0.6518** for LGBM, and an accuracy of **0.9298** for CatBoost.

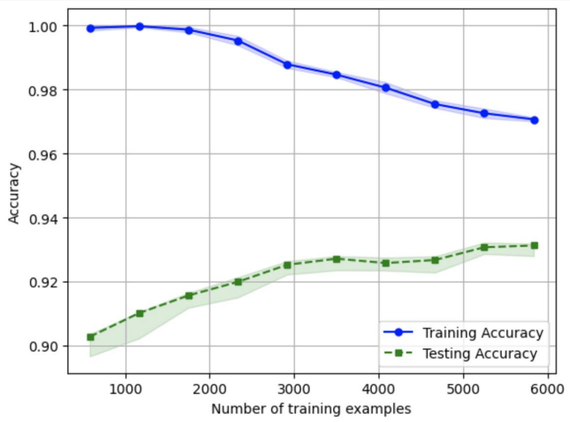
XGBoost



CatBoost



LightGBM



The 3 models are combined to create a well-rounded stacked model.

Stacking / Ensemble

Algorithm \ Metric	Recall	F1 Score	ROC-AUC	Accuracy
CatBoost (Default parameters)	0.560	0.659	0.771	0.934
LightGBM (Default parameters)	0.563	0.654	0.771	0.932
Stacked Model	0.529	0.641	0.757	0.933
Ensemble Model	0.560	0.663	0.772	0.935

		Predict	
		0	1
Actual	0	1272	20
	1	77	89

03.

Contributions to the business

Main areas of contribution

Pricing Segmentation

- **Adverse selection:** When individuals with a higher risk of making a claim are more likely to purchase insurance, while those with lower risk may opt out
- The main goal of insurance pricing is to **segment** risk effectively and avoid adverse selection
- Leveraging policy data and ML models enables insurers to develop more refined pricing structures
- This leads to several benefits:

Increases profitability

Minimizing high-risk exposure may lead to increased profitability

Improves customer retention

Lower premiums for low-risk individuals will improve retention of desired customers

Increases competitiveness

Lower premiums can contribute to a more competitive position for the insurer

Underwriting performance

- Underwriters assess variables such as age, gender and education to decide whether an individual is insurable
- ML models can improve risk classification accuracy, reducing high-risk policies and boosting profitability for insurers

Traditional methods	Machine learning methods
Considers limited number of variables	Can consider larger number of variables
High explainability	Lower explainability
Standard predictability	High predictability

KPI Monitoring

<u>Risk-adjusted Lifetime Value</u>		MSLC	
		High	Low
CLV	High	Desired	Monitor average CLV
	Low	Monitor customer history	Undesired

Monitor metrics for business goals - customer retention and profitability considerations.

Potential Risks & Concerns

Privacy concerns

- Only collect and store information necessary for the construction and maintaining of the model
- Limit access to the database adequately to reduce risk of a major data breach

Fairness and bias

- Establish a system for frequent and adequate maintenance of the model
- Follow the E-23 Model Risk Management guideline to ensure that the model follows industry standards

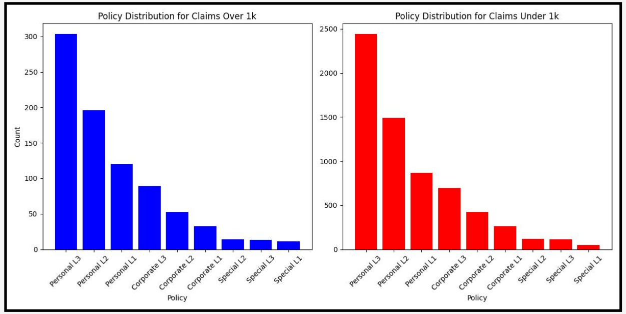
Explainability

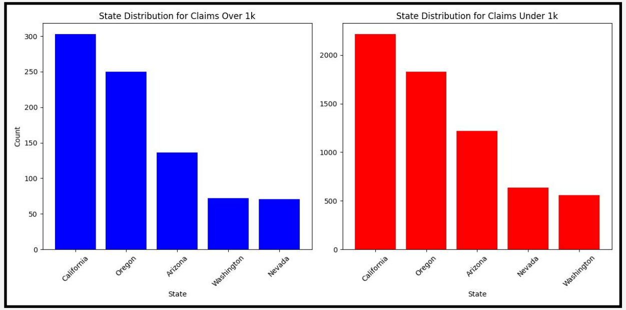
- Analyze SHAP values to see the overall importance of each variable
- Utilize graphs to visualize individual trees within the model

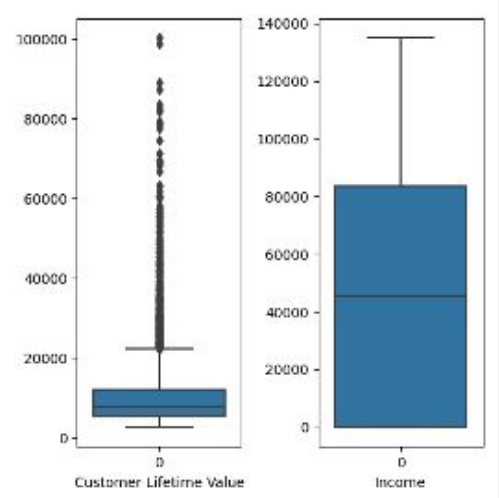
Thank you!

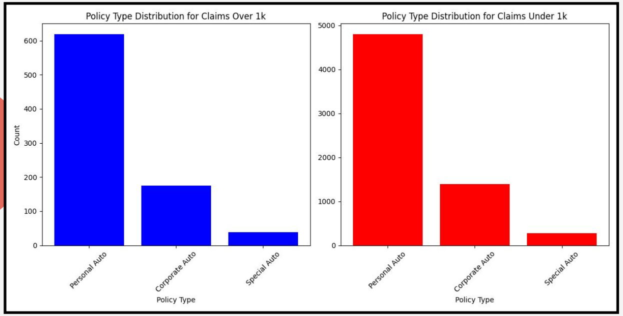
Any questions?

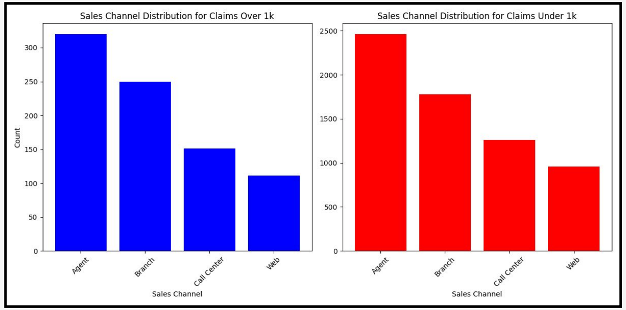
Appendix


Policy Distribution for Claims Over 1k
Policy Distribution for Claims Under 1k
Count
Personal L3
Personal L2
Personal L1
Corporate L3
Corporate L2
Corporate L1
Special L2
Special L3
Special L1
Policy


State Distribution for Claims Over 1k
State Distribution for Claims Under 1k
Count
California
Oregon
Arizona
Washington
Nevada
State

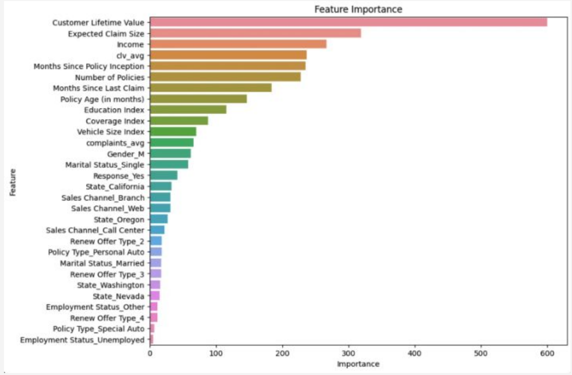

Customer Lifetime Value
Income


Policy Type Distribution for Claims Over 1k
Policy Type Distribution for Claims Under 1k
Count
Personal Auto
Corporate Auto
Special Auto
Policy Type

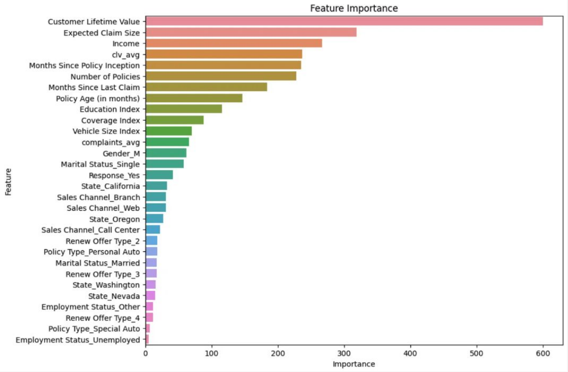

Sales Channel Distribution for Claims Over 1k
Sales Channel Distribution for Claims Under 1k
Count
Agent
Branch
Call Center
Web
Sales Channel

"State", "Policy", "Sales Channel", "Renew Offer Type", "Policy Type"

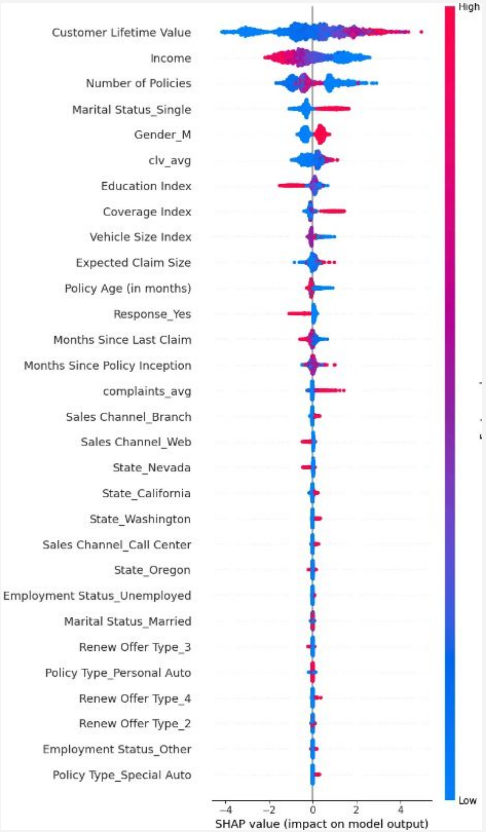
LGBM



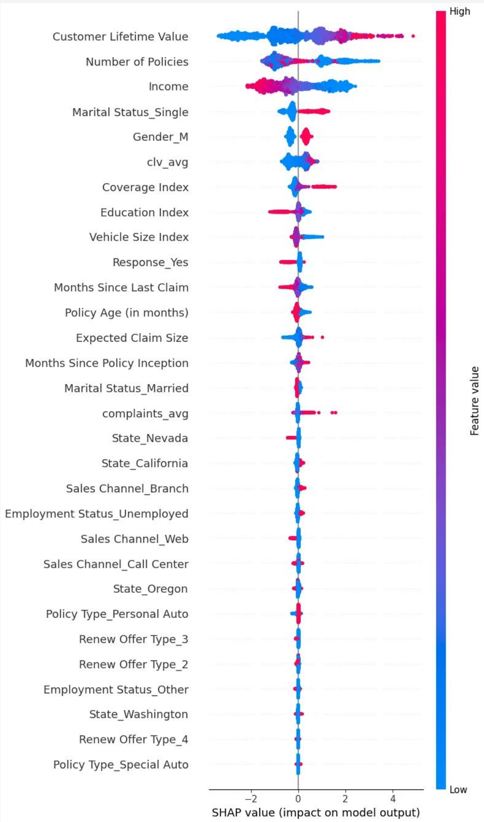
CatBoost



LGBM



CatBoost



Appendix

Model's performance after feature selection

Algorithm \ Metric	**Recall**	**F1 Score**	**ROC-AUC**	**Accuracy**
LightGBM	0.5228	0.6196	0.9116	0.9274
CatBoost	0.5275	0.6266	0.9176	0.9288
XGBoost	0.5226	0.6027	0.9005	0.9222

XGBoost

		Predict	
		0	1
Actual	0	629	17
	1	34	49

CatBoost

		Predict	
		0	1
Actual	0	635	11
	1	35	48

LightGBM

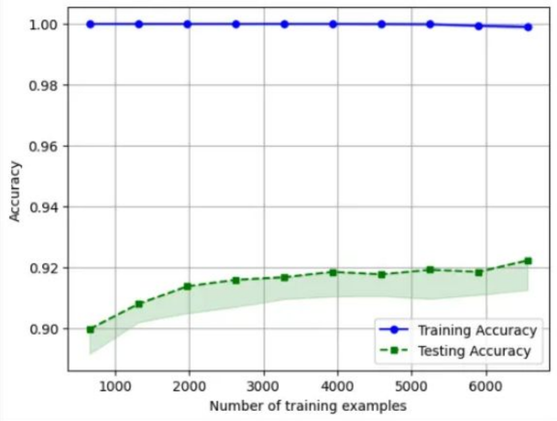
		Predict	
		0	1
Actual	0	632	14
	1	37	46

Appendix

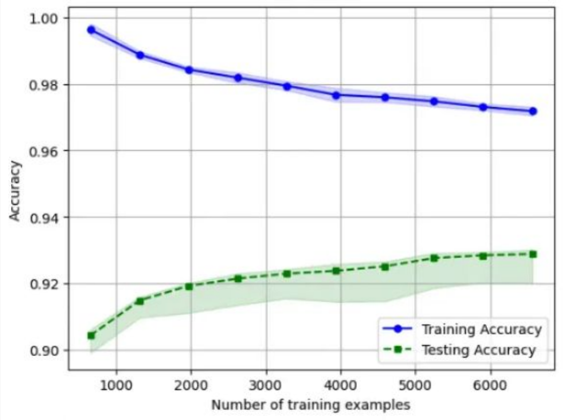
Model's performance after feature selection

Algorithm \ Metric	**Recall**	**F1 Score**	**ROC-AUC**	**Accuracy**
LightGBM	0.5228	0.6196	0.9116	0.9274
CatBoost	0.5275	0.6266	0.9176	0.9288
XGBoost	0.5226	0.6027	0.9005	0.9222

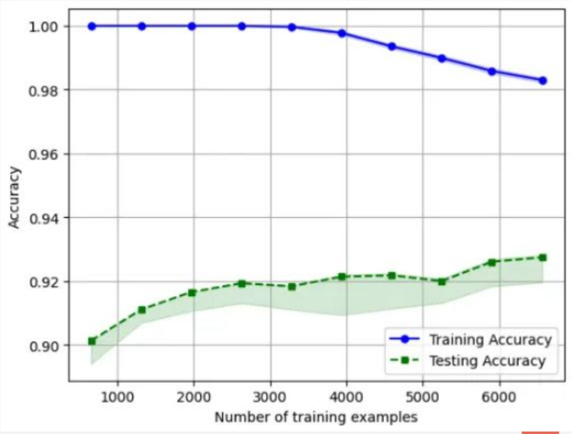
XGBoost



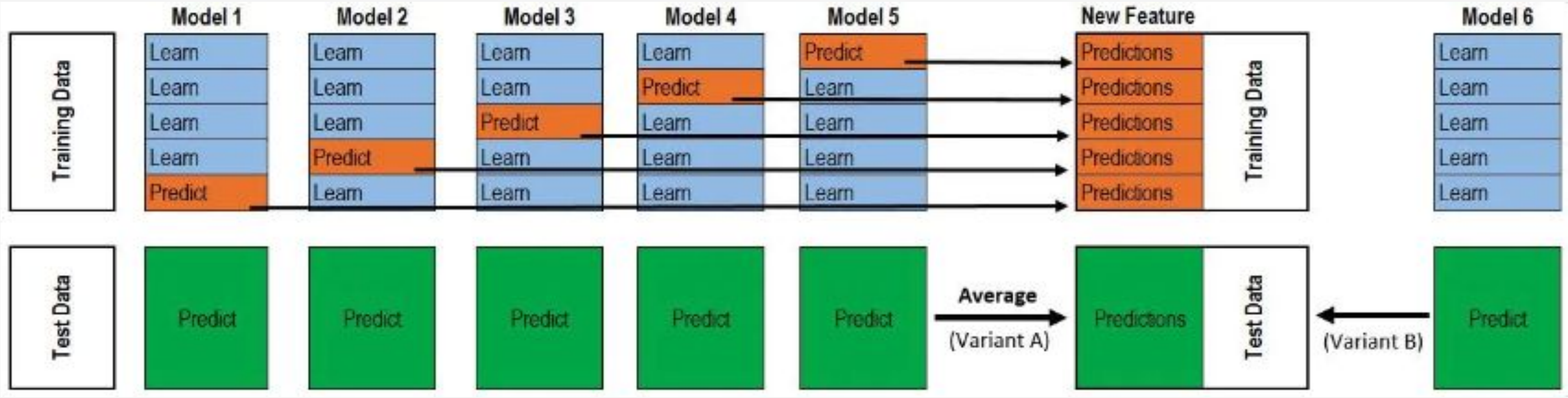
CatBoost



LightGBM



Appendix



Model 1
Model 2
Model 3
Model 4
Model 5
New Feature
Model 6
Training Data
Learn
Predict
Predictions
Training Data
Test Data
Predict
Average
(Variant A)
Predictions
Test Data
(Variant B)
Predict