ML Raport

AutoPrep

January 11, 2025

Abstract

This raport has been generated with AutoPrep.

Contents

1 Overview

1.1 System

System	Darwin
Machine	arm64
Processor	arm
Architecture	64bit
Python Version	3.10.5
Physical Cores	8
Logical Cores	8
CPU Frequency (MHz)	3204
Total RAM (GB)	16.00
Available RAM (GB)	4.44
Total Disk Space (GB)	228.27
Free Disk Space (GB)	7.20

Table 1: System overview.

1.2 Dataset

Task detected for the dataset: multiclass classfication.
Table 2 presents an overview of the dataset including the number of samples, features, and their types.

Number of samples	124
Number of features	8
Number of numerical features	2
Number of categorical features	6

Table 2: Dataset Summary.

Distribution of the target classes in terms of the number of observations and their percentages is presented in Table 3

class	**number of observations**	**fraction**
low	40	0.32
high	34	0.27
average	32	0.26
veryhigh	18	0.15

Table 3: Target class distribution.

Table 4 presents the distribution of missing values in the dataset.

feature	number of observations	fraction
year_zone	0	0.00
year	0	0.00
strip	0	0.00
pdk	0	0.00
damage_rankRJT	0	0.00
damage_rankALL	0	0.00
dry_or_irr	0	0.00
zone	0	0.00

Table 4: Missing values distribution.

Table 5 presents the description of features in the dataset.

feature	type	dtype	space usage
year_zone	categorical	category	2.9 kB
year	categorical	category	1.8 kB
strip	numerical	uint8	1.1 kB
pdk	numerical	uint8	1.1 kB
damage_rankRJT	categorical	category	1.6 kB
damage_rankALL	categorical	category	1.6 kB
dry_or_irr	categorical	category	1.4 kB
zone	categorical	category	1.4 kB

Table 5: Features dtypes description.

Table 6 and Table 7 present the description of numerical and categorical features in the dataset.

feature	count	mean	std	min	25%	50%	75%	max
strip	124.00	5.24	3.16	1.00	3.00	5.00	9.00	10.00
pdk	124.00	2.23	1.06	0.00	1.00	2.00	3.00	5.00

Table 6: Numerical features description.

index	count	unique	top	freq
year_zone	124	21	9f	11
year	124	7	91	22
damage_rankRJT	124	6	1	31
damage_rankALL	124	6	1	36
dry_or_irr	124	3	D	102
zone	124	3	F	61

Table 7: Categorical features description.

2 Eda

This part of the report provides basic insides to the data and the informations it holds..

2.1 Target variable and missing values

Figure 1 shows the distribution of the target variable.

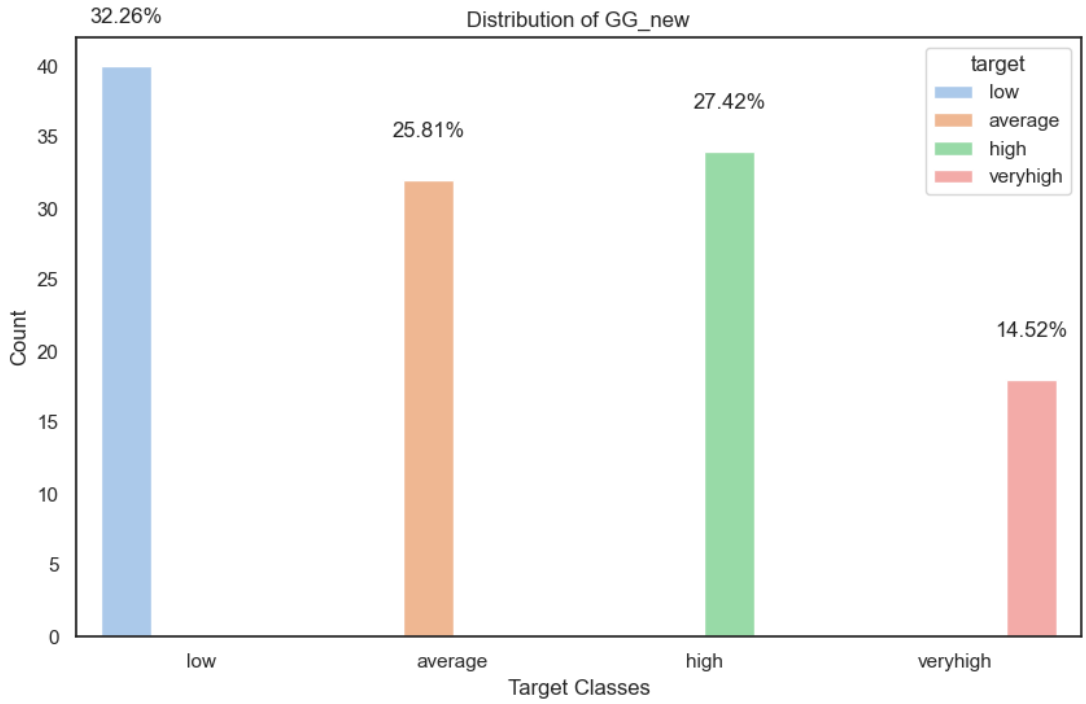

Figure 1: Target distribution.

2.2 EDA for categorical features

The distribution of categorical features is presented on barplot(s) below.

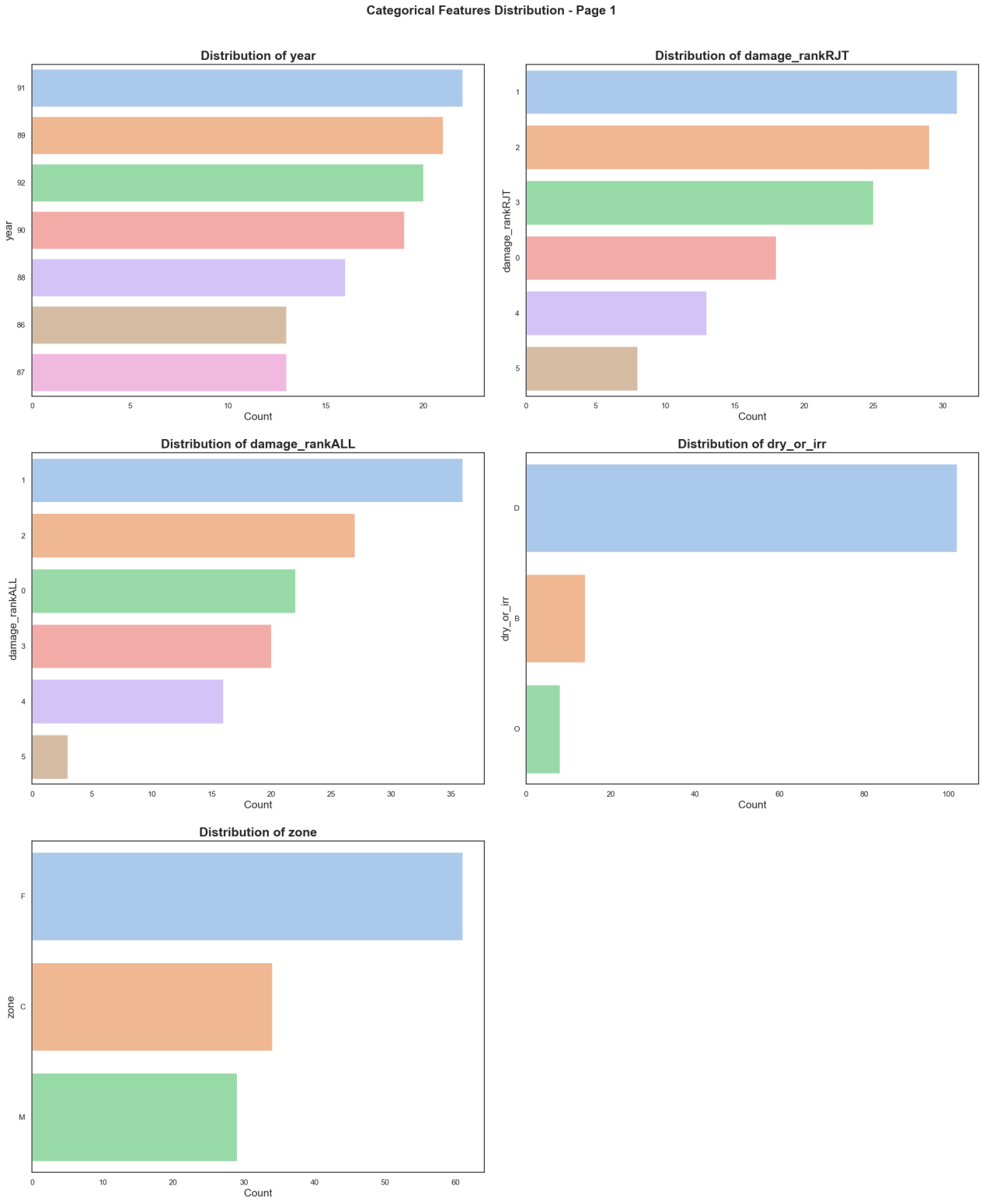

Figure 2: Categorical Features Distribution - Page 1

2.3 EDA for numerical features

The distribution of numerical features is presented on histogram(s) below.

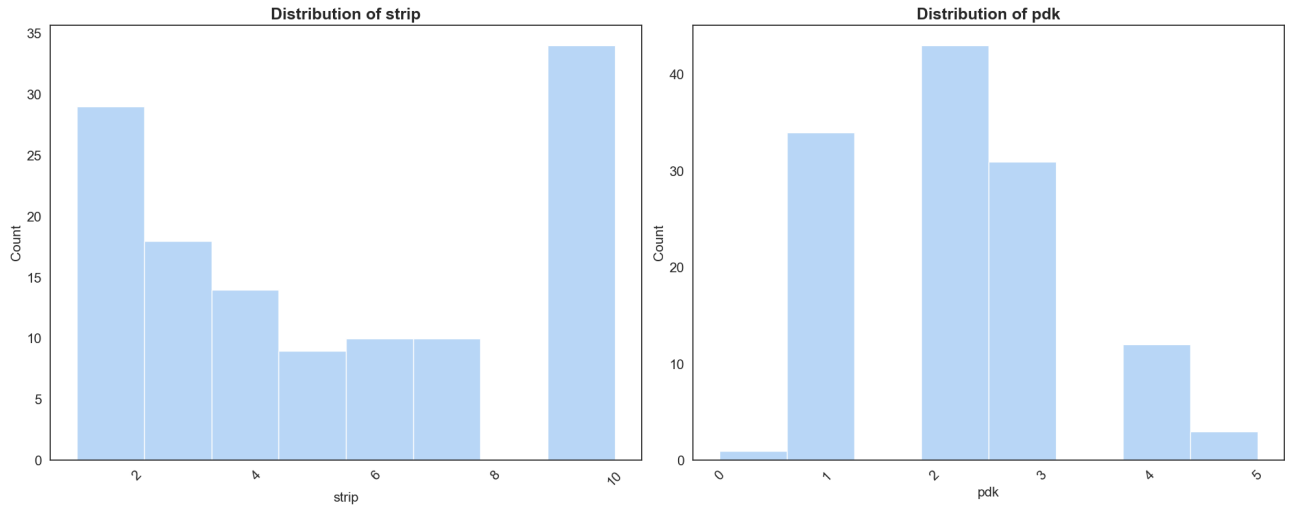

Figure 3: Numerical Features Distribution - Page 1

Figure 4 shows the correlation between features.

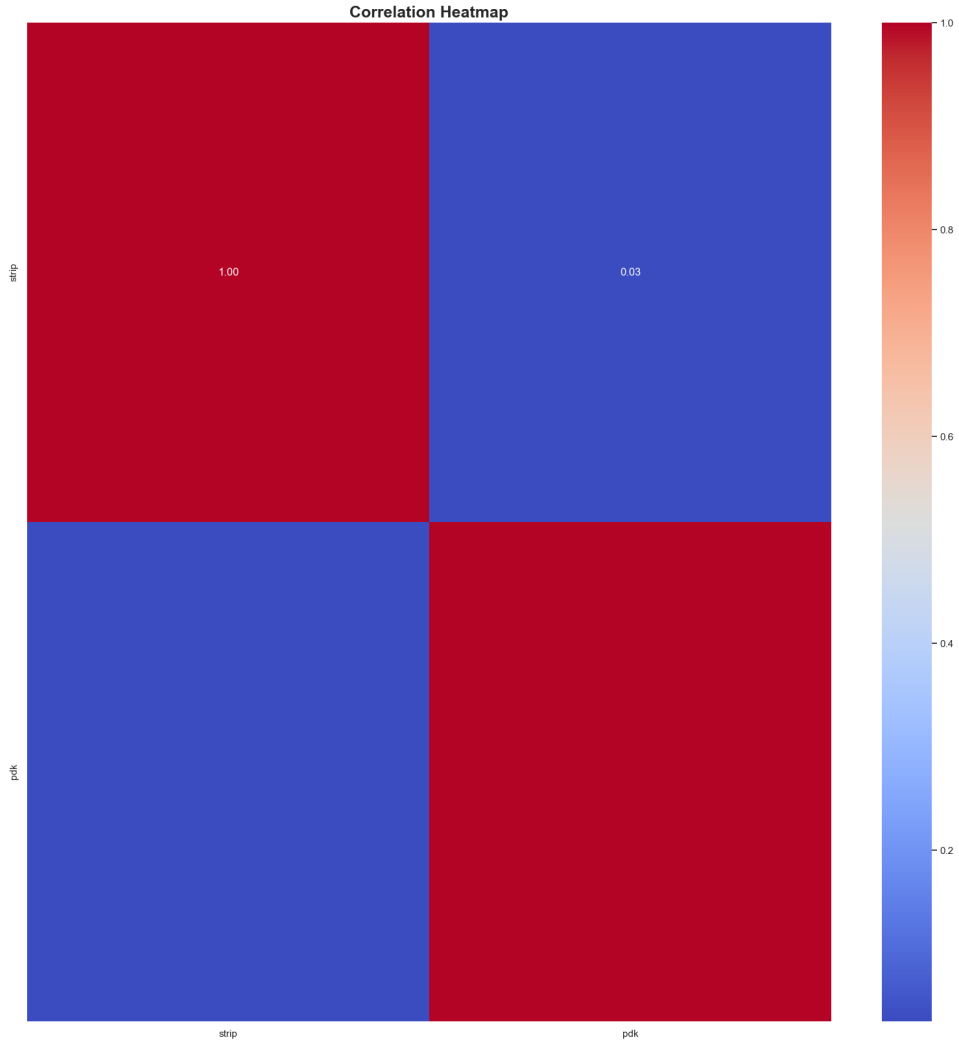

Figure 4: Correlation heatmap.

The boxplot of numerical features is presented on chart(s) below.

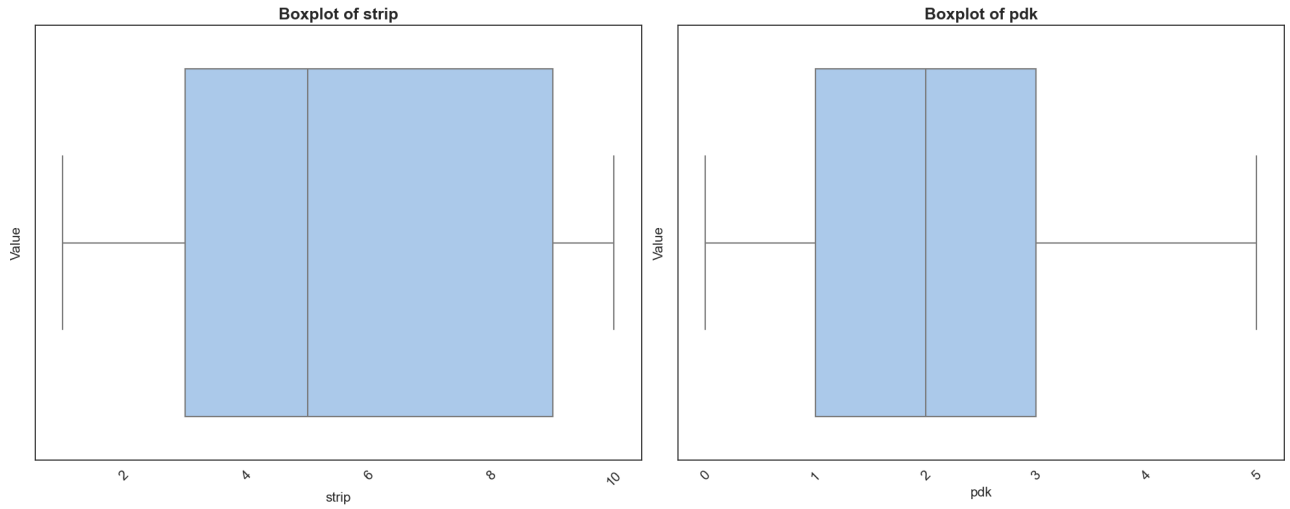

Figure 5: Boxplot page 1

3 Preprocessing

This part of the report presents the results of the preprocessing process. It contains required, as well as non required, steps listed below.
Required preprocessing steps

- Missing data imputation
- Removing columns with 100% unique categorical values
- Categorical features encoding
- Scaling
- Removing columns with 0 variance
- Detecting highly correlatd features

Additional preprocessing steps

- Feature selection methods : Correlation with the target or Random Forest feature importance
- Dimention reduction techniques: PCA, VIF, UMAP

Preprocessing process was configured to select up to 3 best unique preprocessing pipelines. Pipelines were scored based on a simple model. Tables below show detailed description of the best pipelines as well as all step combinations that were examined.

index	steps
0	NAImputer, UniqueFilter, ColumnEncoder, VarianceFilter, CorrelationFilter, ColumnScaler

Table 8: Pipelines steps overview.

index	file name	score	fit duration	score duration
0	preprocessing_pipeline_0.joblib	0.67	a moment	a moment
1	preprocessing_pipeline_1.joblib	0.67	a moment	a moment
2	preprocessing_pipeline_2.joblib	0.67	a moment	a moment

Table 9: Best preprocessing pipelines.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}
4	CorrelationFilter	Removes one column from pairs of columns correlated above correlation threshold: 0.8.	{}
5	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "standard"}

Table 10: Best pipeline No. 0: steps overview.

index	count	mean	std	min	25%	50%	75%	max
year_zone	124.00	0.00	1.00	-1.53	-0.89	-0.10	0.89	1.64
year	124.00	-0.00	1.00	-1.74	-0.70	-0.18	0.86	1.39
strip	124.00	-0.00	1.00	-1.35	-0.71	-0.08	1.19	1.51
pdk	124.00	0.00	1.00	-2.11	-1.16	-0.21	0.73	2.63
damage_rankRJT	124.00	-0.00	1.00	-1.45	-0.75	-0.05	0.66	2.06
damage_rankALL	124.00	0.00	1.00	-1.35	-0.62	0.11	0.84	2.30
dry_or_irr_B	124.00	0.00	1.00	-0.36	-0.36	-0.36	-0.36	2.80
dry_or_irr_D	124.00	-0.00	1.00	-2.15	0.46	0.46	0.46	0.46
dry_or_irr_O	124.00	-0.00	1.00	-0.26	-0.26	-0.26	-0.26	3.81
zone_M	124.00	-0.00	1.00	-0.55	-0.55	-0.55	-0.55	1.81

Table 11: Best pipeline No. 0: Output overview.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}
4	CorrelationFilter	Removes one column from pairs of columns correlated above correlation threshold: 0.8.	{}
5	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "minmax"}

Table 12: Best pipeline No. 1: steps overview.

index	count	mean	std	min	25%	50%	75%	max
year_zone	124.00	0.48	0.32	0.00	0.20	0.45	0.76	1.00
year	124.00	0.56	0.32	0.00	0.33	0.50	0.83	1.00
strip	124.00	0.47	0.35	0.00	0.22	0.44	0.89	1.00
pdk	124.00	0.45	0.21	0.00	0.20	0.40	0.60	1.00
damage_rankRJT	124.00	0.41	0.29	0.00	0.20	0.40	0.60	1.00
damage_rankALL	124.00	0.37	0.27	0.00	0.20	0.40	0.60	1.00
dry_or_irr_B	124.00	0.11	0.32	0.00	0.00	0.00	0.00	1.00
dry_or_irr_D	124.00	0.82	0.38	0.00	1.00	1.00	1.00	1.00
dry_or_irr_O	124.00	0.06	0.25	0.00	0.00	0.00	0.00	1.00
zone_M	124.00	0.23	0.43	0.00	0.00	0.00	0.00	1.00

Table 13: Best pipeline No. 1: Output overview.

step	name	description	params
0	NAImputer	Imputes missing data.	{"numeric_imputer": "median", "categorical_imputer": "most_frequent"}
1	UniqueFilter	Removes categorical columns with 100% unique values. Dropped columns: []	{}
2	ColumnEncoder	Encodes categorical columns using OneHotEncoder (for columns with <5 unique values) or TolerantLabelEncoder (for columns with >=5 unique values). Encodes target variable using LabelEncoder if provided.	{}
3	VarianceFilter	Removes columns with zero variance. Dropped columns: []	{}
4	CorrelationFilter	Removes one column from pairs of columns correlated above correlation threshold: 0.8.	{}
5	ColumnScaler	Scales numerical columns using one of 3 scaling methods.	{"method": "robust"}

Table 14: Best pipeline No. 2: steps overview.

index	count	mean	std	min	25%	50%	75%	max
year_zone	124.00	0.06	0.56	-0.80	-0.44	0.00	0.56	0.98
year	124.00	0.11	0.64	-1.00	-0.33	0.00	0.67	1.00
strip	124.00	0.04	0.53	-0.67	-0.33	0.00	0.67	0.83
pdk	124.00	0.11	0.53	-1.00	-0.50	0.00	0.50	1.50
damage_rankRJT	124.00	0.03	0.71	-1.00	-0.50	0.00	0.50	1.50
damage_rankALL	124.00	-0.08	0.69	-1.00	-0.50	0.00	0.50	1.50
dry_or_irr_B	124.00	0.11	0.32	0.00	0.00	0.00	0.00	1.00
dry_or_irr_D	124.00	-0.18	0.38	-1.00	0.00	0.00	0.00	0.00
dry_or_irr_O	124.00	0.06	0.25	0.00	0.00	0.00	0.00	1.00
zone_M	124.00	0.23	0.43	0.00	0.00	0.00	0.00	1.00

Table 15: Best pipeline No. 2: Output overview.

You may also find all pipelines' runtime statistic in Table 16

Category	Value
Unique created pipelines	1
All created pipelines (after exploading each step params)	3
All pipelines fit time	2 seconds
All pipelines score time	3 seconds
scores_count	3.00
scores_mean	0.67
scores_std	0.00
scores_min	0.67
scores_25%	0.67
scores_50%	0.67
scores_75%	0.67
scores_max	0.67
Scoring function	<class 'str'>
Scoring model	RandomForestClassifier

Table 16: Preprocessing pipelines runtime statistics.

4 Modeling

4.1 Overview

This part of the report presents the results of the modeling process. There were 5 classification models trained for each of the best preprocessing pipelines.
The following models were used in the modeling process.

- KNeighborsClassifier
- LogisticRegression
- GaussianNB
- SVC
- DecisionTreeClassifier

4.2 Hyperparameter tuning

This section presents the results of hyperparameter tuning for each of the best 3 models using RandomizedSearchCV. Param grids used for each model are presented in the tables below.

Category	Value
n_neighbors	[5, 10, 15]
weights	['uniform', 'distance']
algorithm	['auto', 'ball_tree', 'kd_tree', 'brute']
leaf_size	[30, 40, 50]
p	[1, 2]

Table 17: Param grid for model KNeighboursClassifier.

Category	Value
0	{"penalty": ["l1"], "C": [0.01, 0.1, 1, 10], "solver": ["liblinear", "saga"]}
1	{"penalty": ["l2"], "C": [0.01, 0.1, 1, 10], "solver": ["lbfgs", "liblinear", "saga", "newton-cg"]}
2	{"penalty": ["elasticnet"], "C": [0.01, 0.1, 1, 10], "solver": ["saga"], "l1_ratio": [0.5, 0.7]}

Table 18: Param grid for model LogisticRegression.

Category	Value
priors	[None]
var_smoothing	[1e-09, 1e-07, 1e-05]

Table 19: Param grid for model GaussianNaiveClassifier.

Category	Value
C	[0.1, 1, 10, 100, 1000]
kernel	['linear', 'poly', 'rbf', 'sigmoid']
degree	[3, 4, 5]
gamma	['scale', 'auto']
random_state	[42]

Table 20: Param grid for model SVC.

Category	Value
criterion	['gini', 'entropy']
splitter	['best', 'random']
max_depth	[None, 5, 10, 15, 20]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]
random_state	[42]

Table 21: Param grid for model DecisionTreeClassifier.

Table 22 presents the best models and pipelines along with their hyperparameters, mean fit time, and test score.

Model	Pipeline	Best params	Mean fit time	Test score
KNeighborsClassifier	final_pipeline_0.joblib	{"weights": "distance", "p": 1, "n_neighbors": 15, "leaf_size": 30, "algorithm": "brute"}	a moment	0.00
KNeighborsClassifier	final_pipeline_1.joblib	{"weights": "distance", "p": 2, "n_neighbors": 10, "leaf_size": 40, "algorithm": "auto"}	a moment	0.00
KNeighborsClassifier	final_pipeline_2.joblib	{"weights": "uniform", "p": 2, "n_neighbors": 15, "leaf_size": 30, "algorithm": "kd_tree"}	a moment	0.00

Table 22: Best models results

4.3 Interpretability

This section presents SHAP plots for the best model.

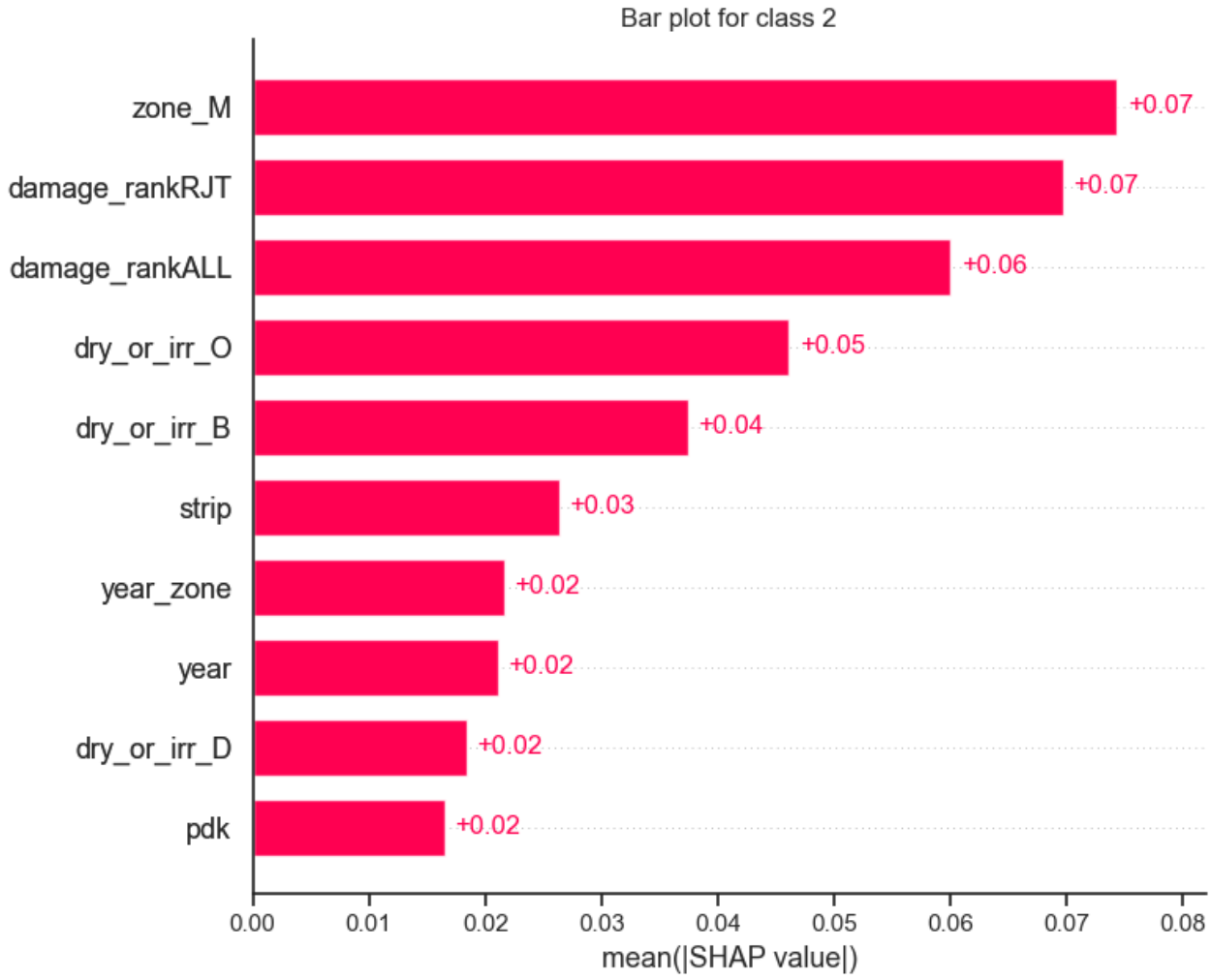

Figure 6: SHAP bar plot for class 2.

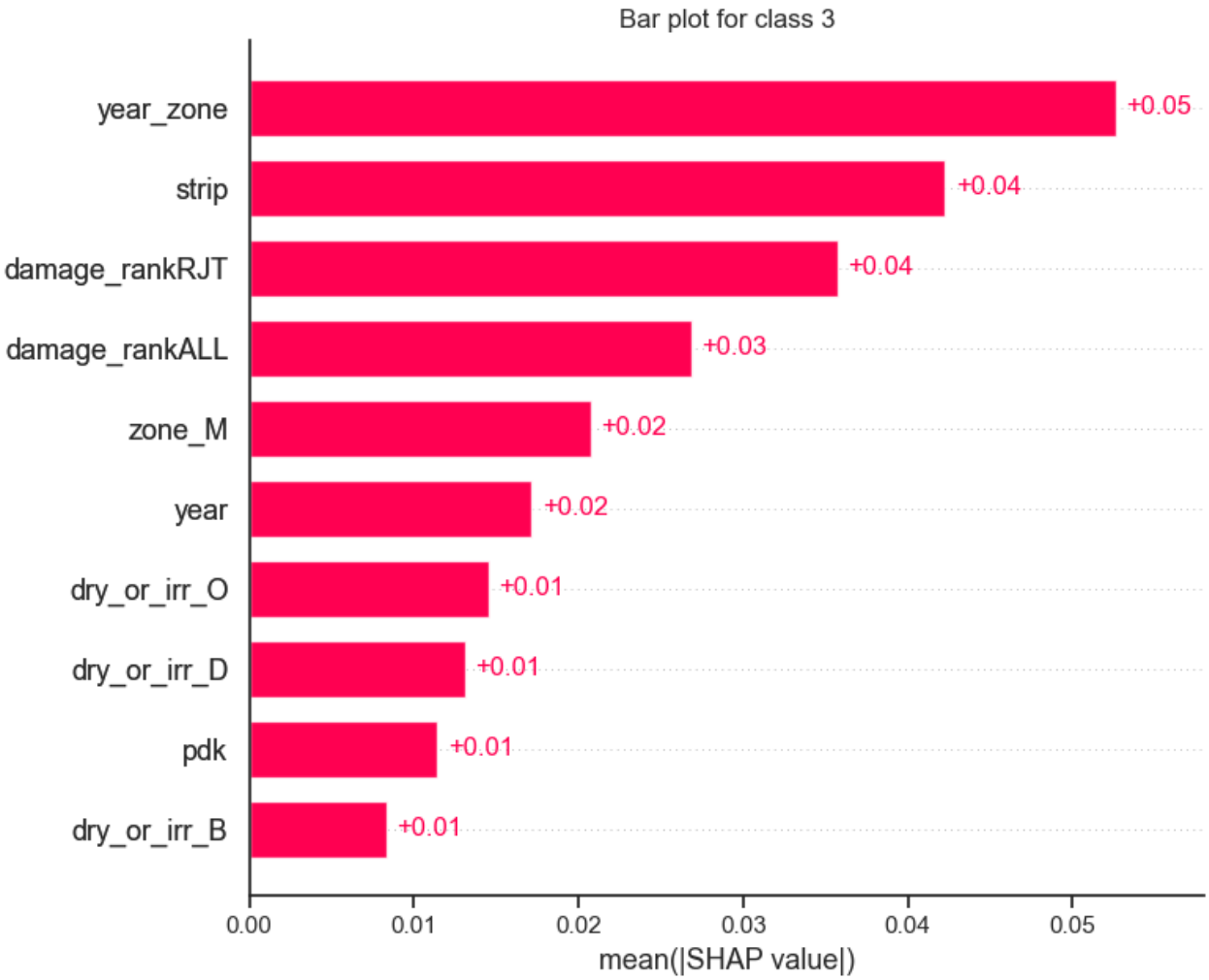

Figure 7: SHAP bar plot for class 3.

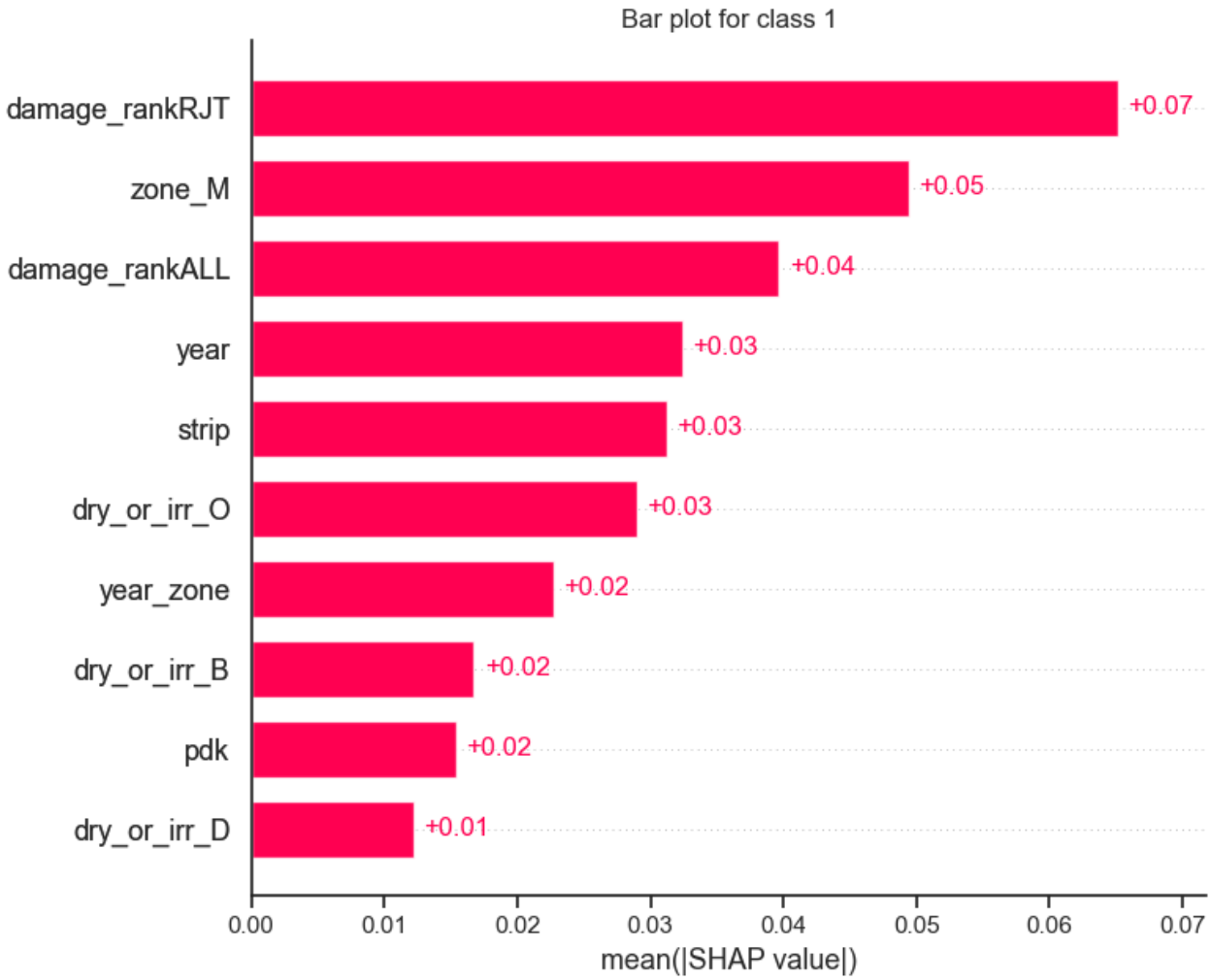

Figure 8: SHAP bar plot for class 1.

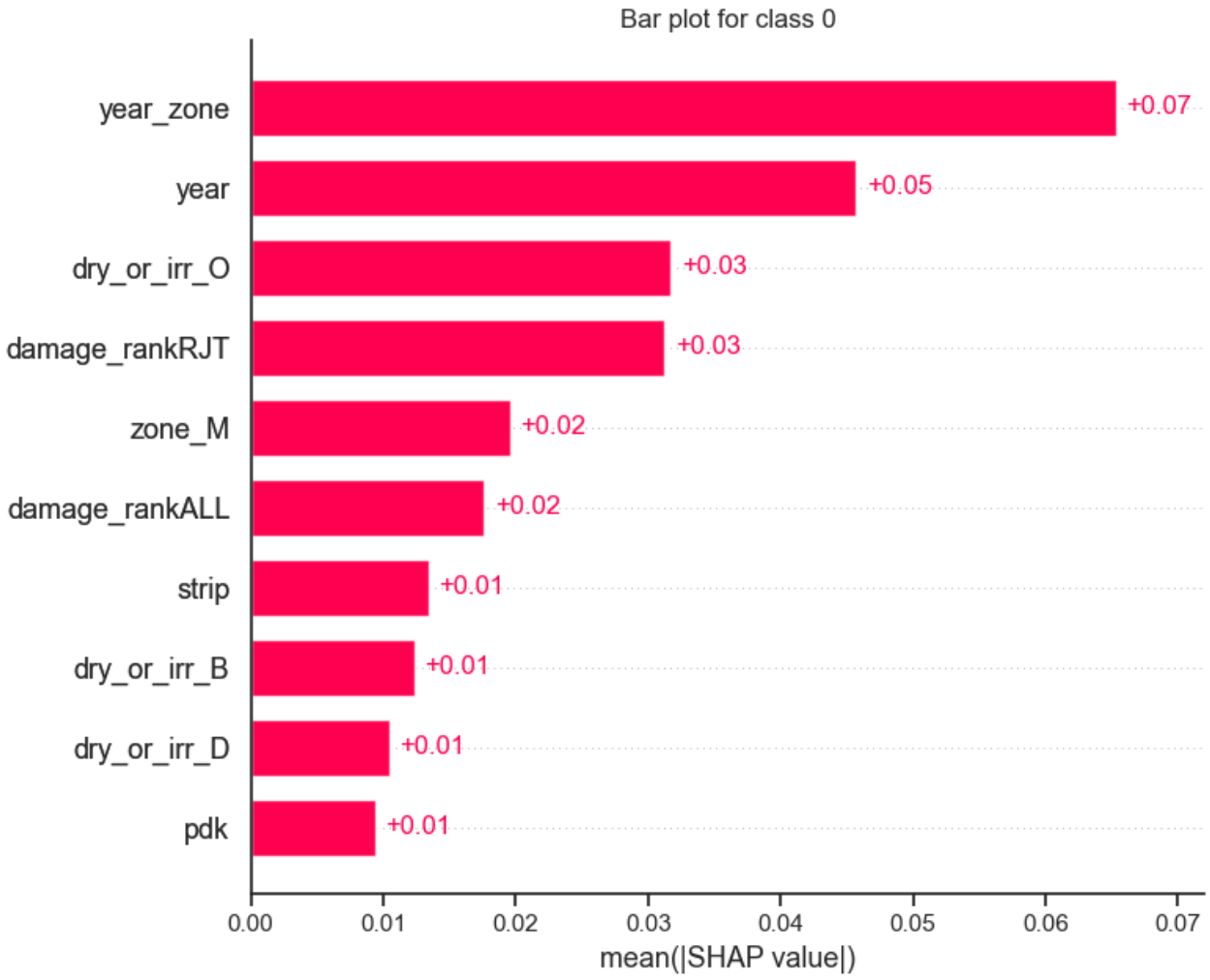

Figure 9: SHAP bar plot for class 0.

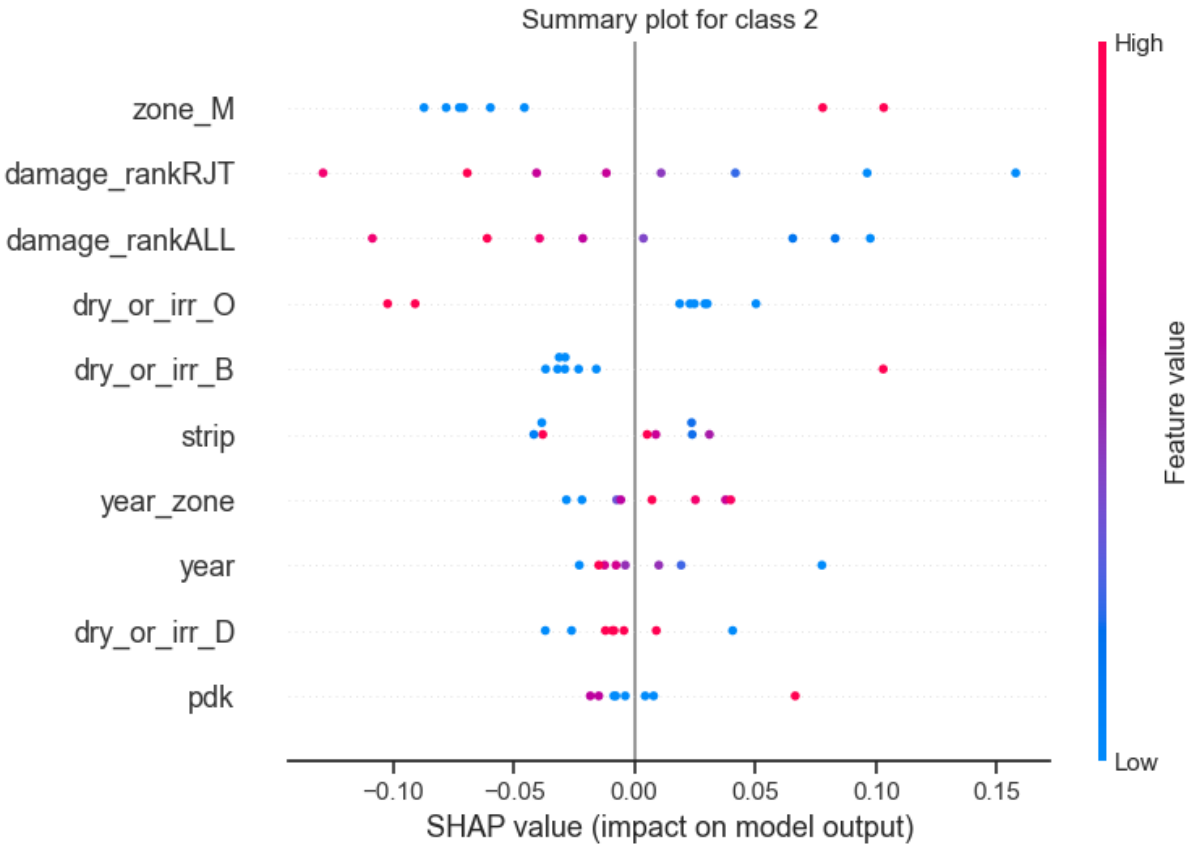

Figure 10: SHAP summary plot for class 2.

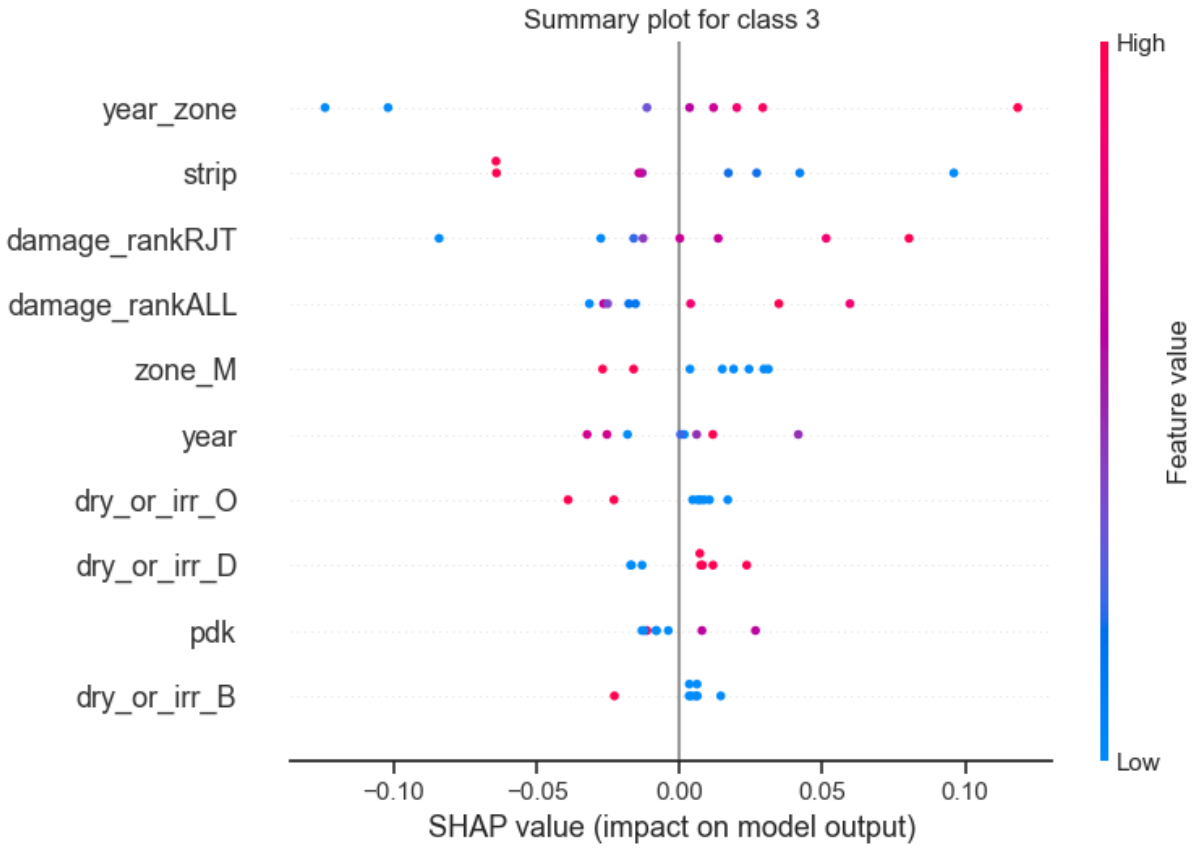

Figure 11: SHAP summary plot for class 3.

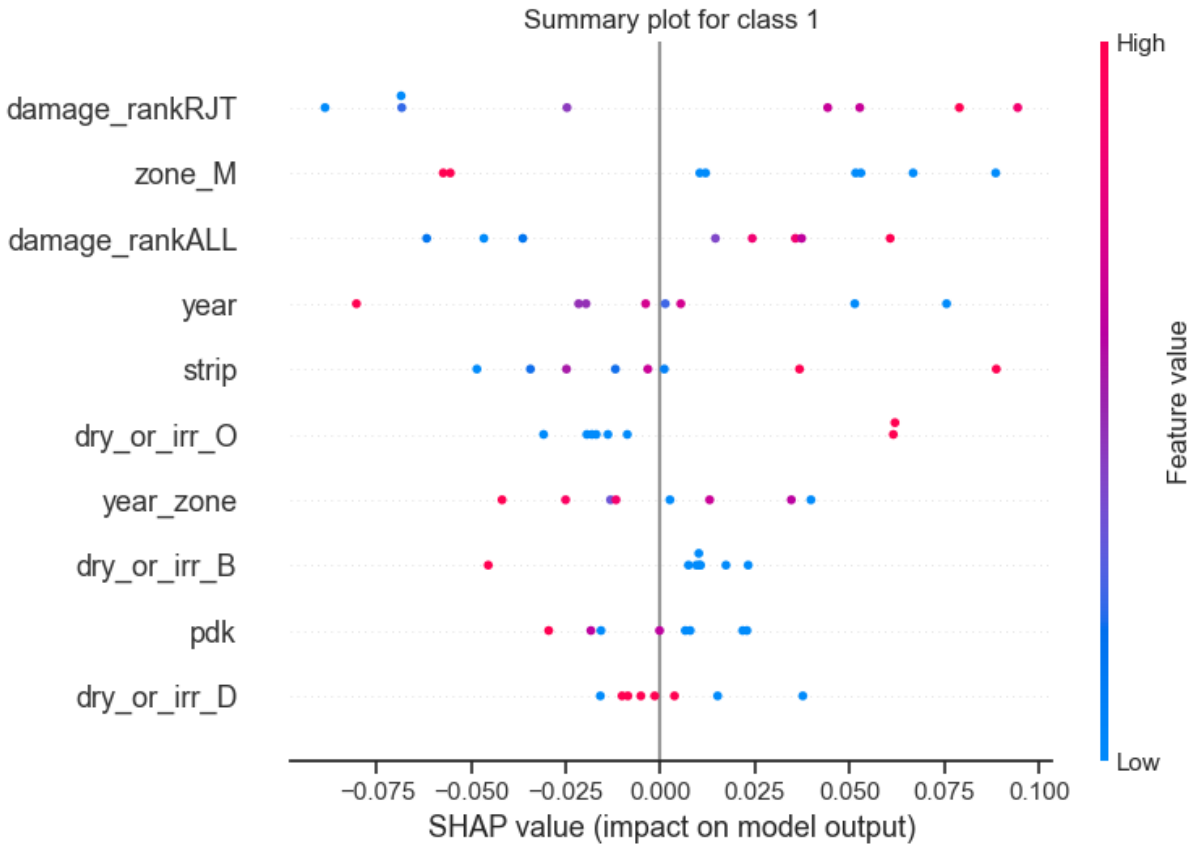

Figure 12: SHAP summary plot for class 1.

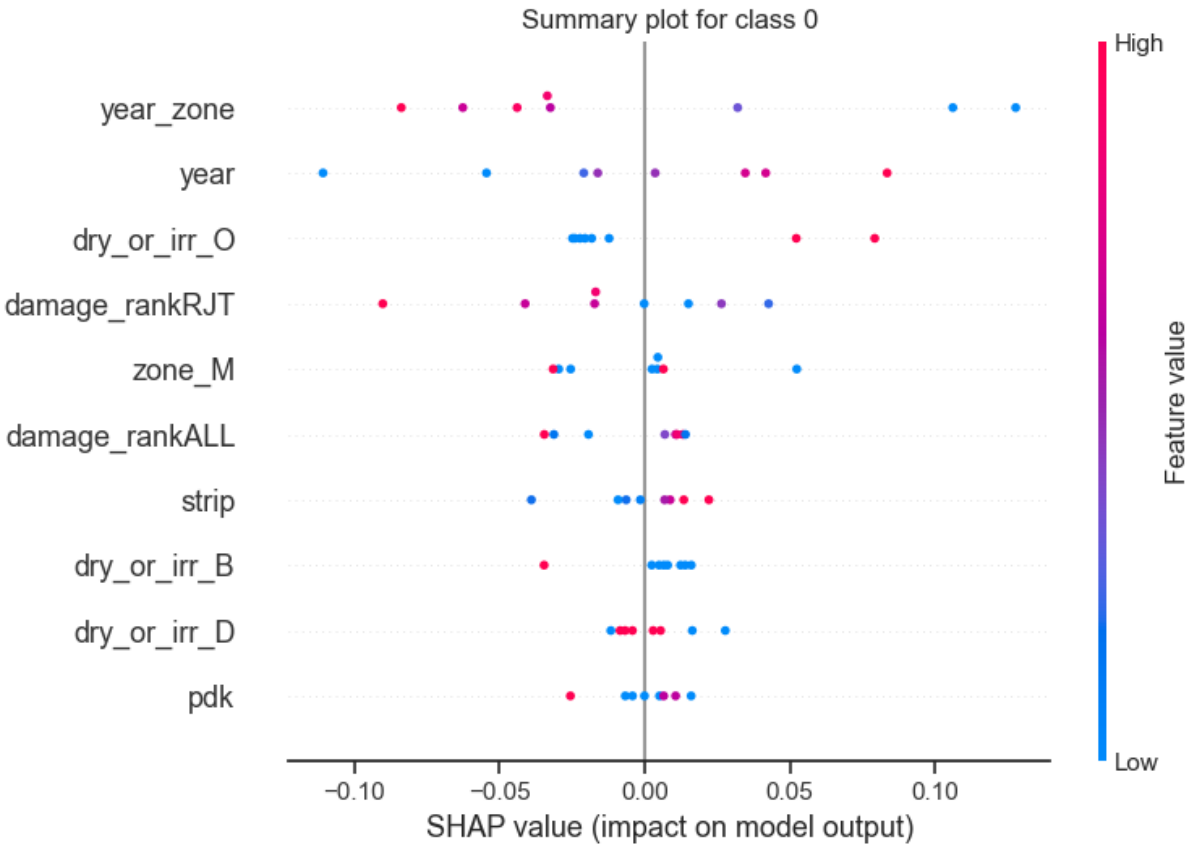

Figure 13: SHAP summary plot for class 0.

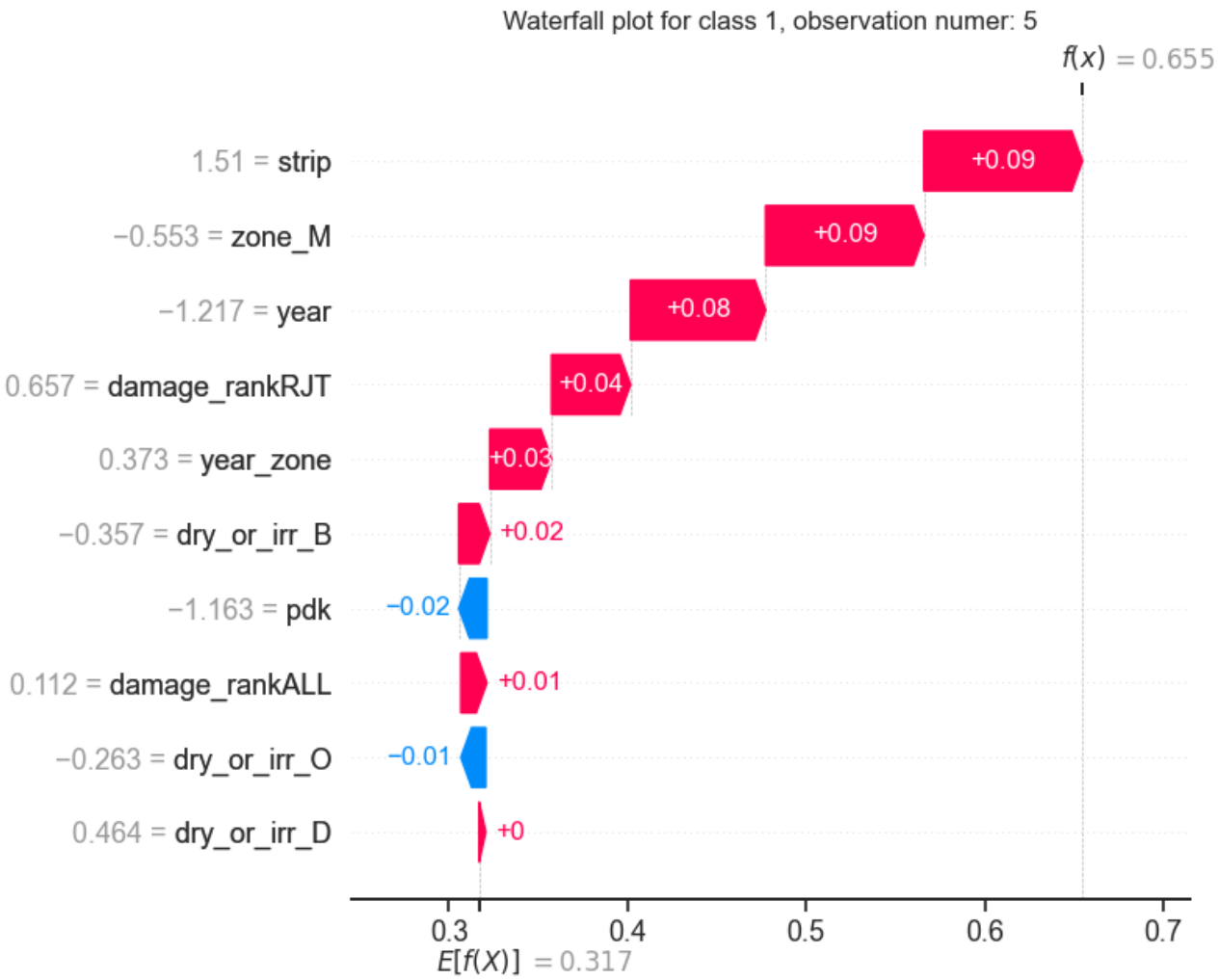

Figure 14: SHAP waterfall plot for class 1.

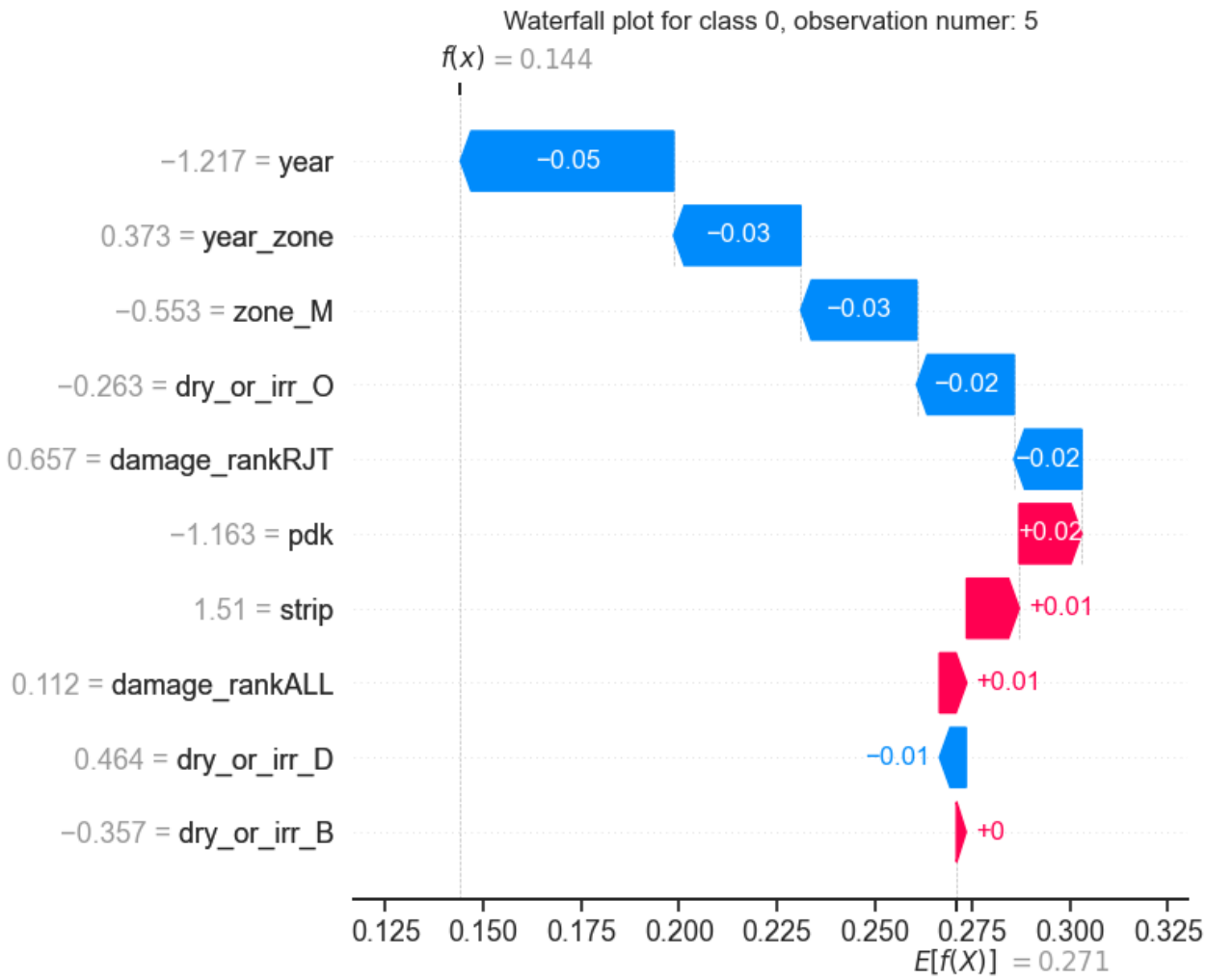

Figure 15: SHAP waterfall plot for class 0.

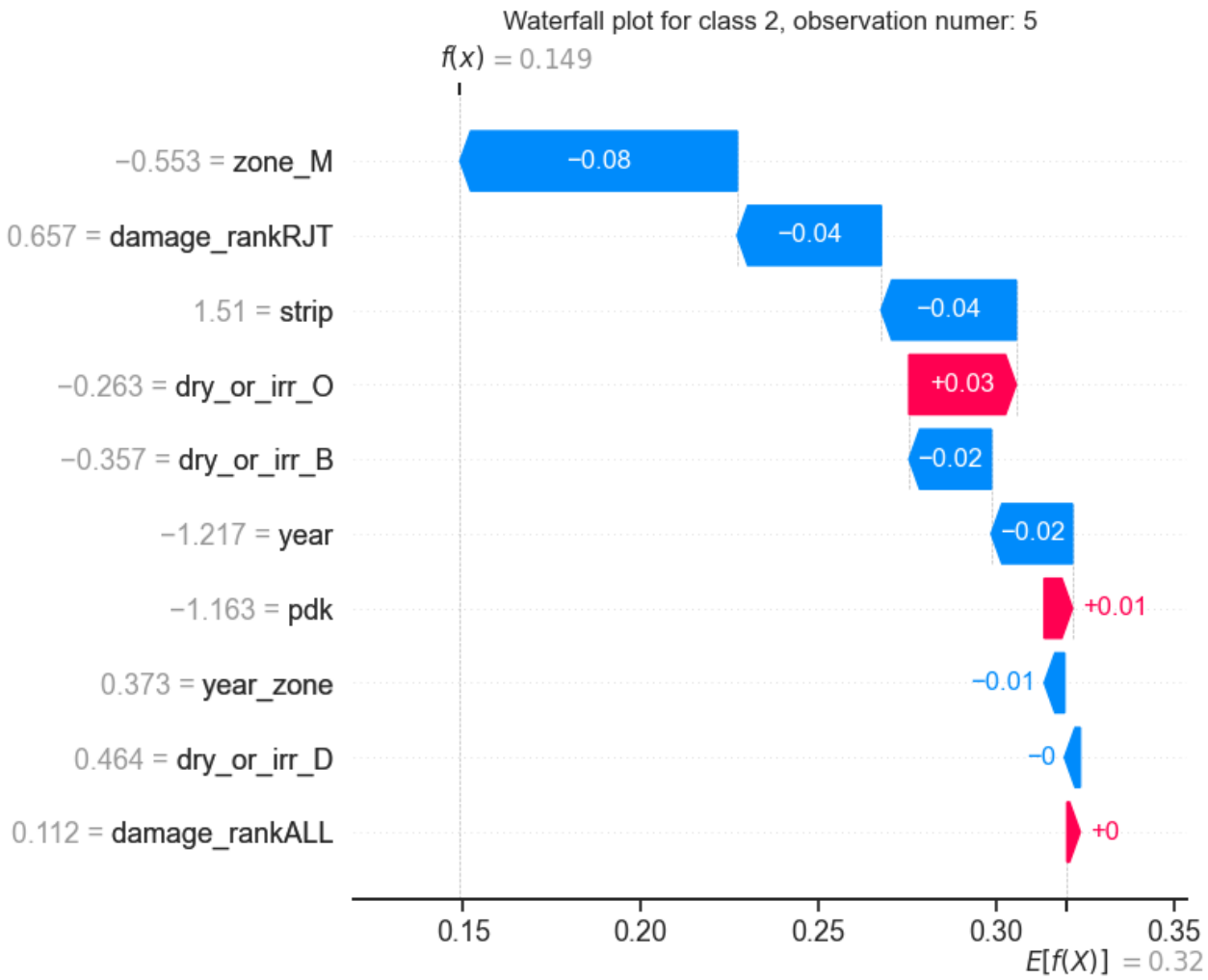

Figure 16: SHAP waterfall plot for class 2.

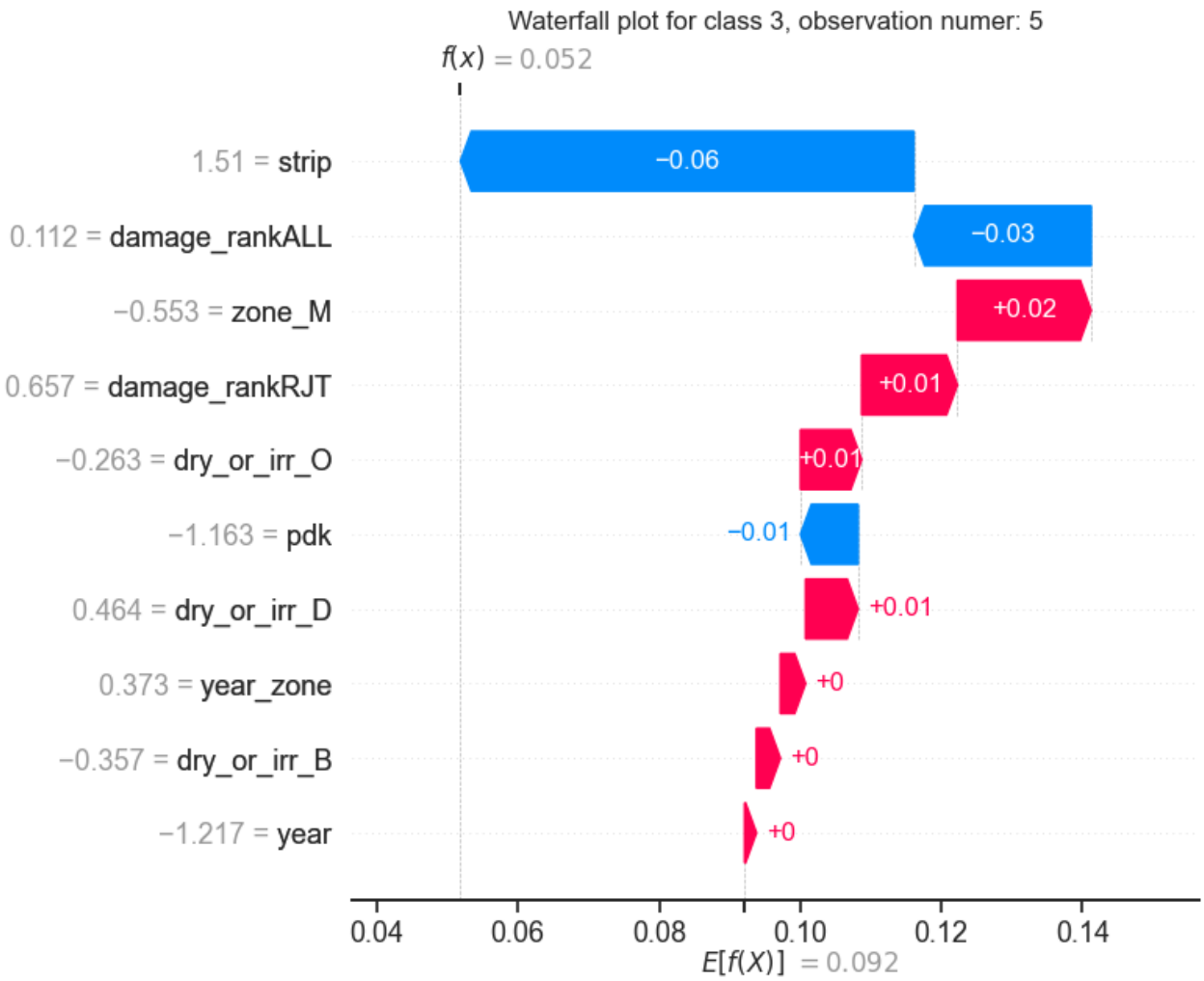

Figure 17: SHAP waterfall plot for class 3.