OREGON FIRE'S CAUSES

JUANITA CORTES, LORELEI GOROCICA AND VICENTE CONSOLI

ML PROJECT

PROBLEM DEFINITION

What are the causes of fires?

Supervised Learning

The model is getting trained on a labelled dataset.

Binary Classification

The **goal** is to classify the fires into classes.

Two classes: caused by human or lighting

DATA COLLECTION

Dataset: ODF Fire Occurrence Data 2000-2022

Oregon Dept of Forestry statistical wildfires from 2000 through 2022. Point locations and fire causes included.

https://data.oregon.gov/Natural-Resources/ODF-Fire-Occurrence-Data-2000-2022/fbwv-q84y

DATA ANALYSIS

PANDAS PROFILING

Understanding all features and their characteristics

DESCRIPTION

Statistical description of all features and dimensions of the dataset.

NULL VALUES

Cleaning data with null values

DISTRIBUTION

Proportion of causes: % of human and lighting

PRESELECTION OF FEATURES

Deleting repetitive features

DATA PREPROCESSING

Label encoding

Encoding target labels with value between 0 and n_classes

Feature selection

Chi - square method

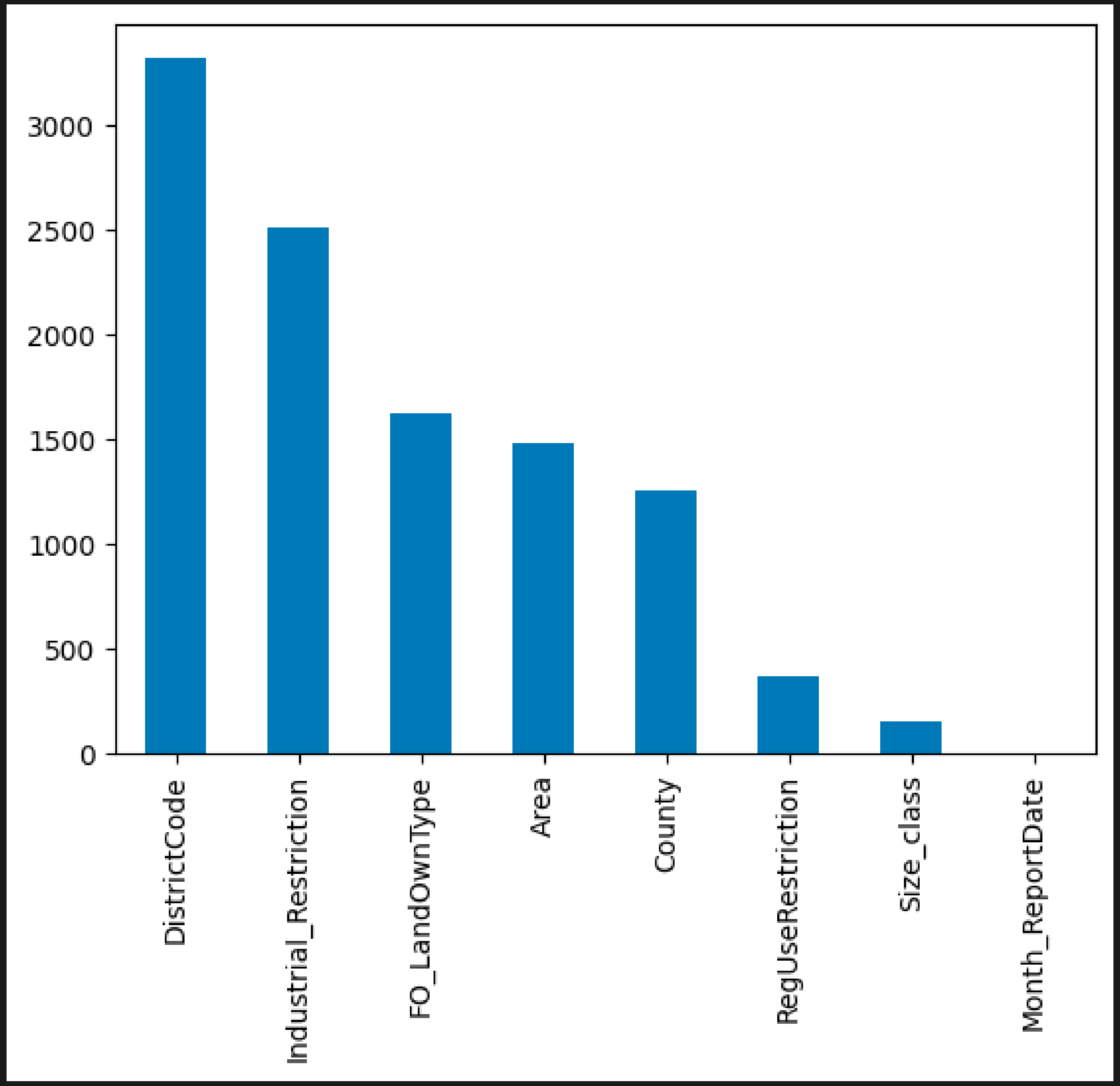
Data spliting

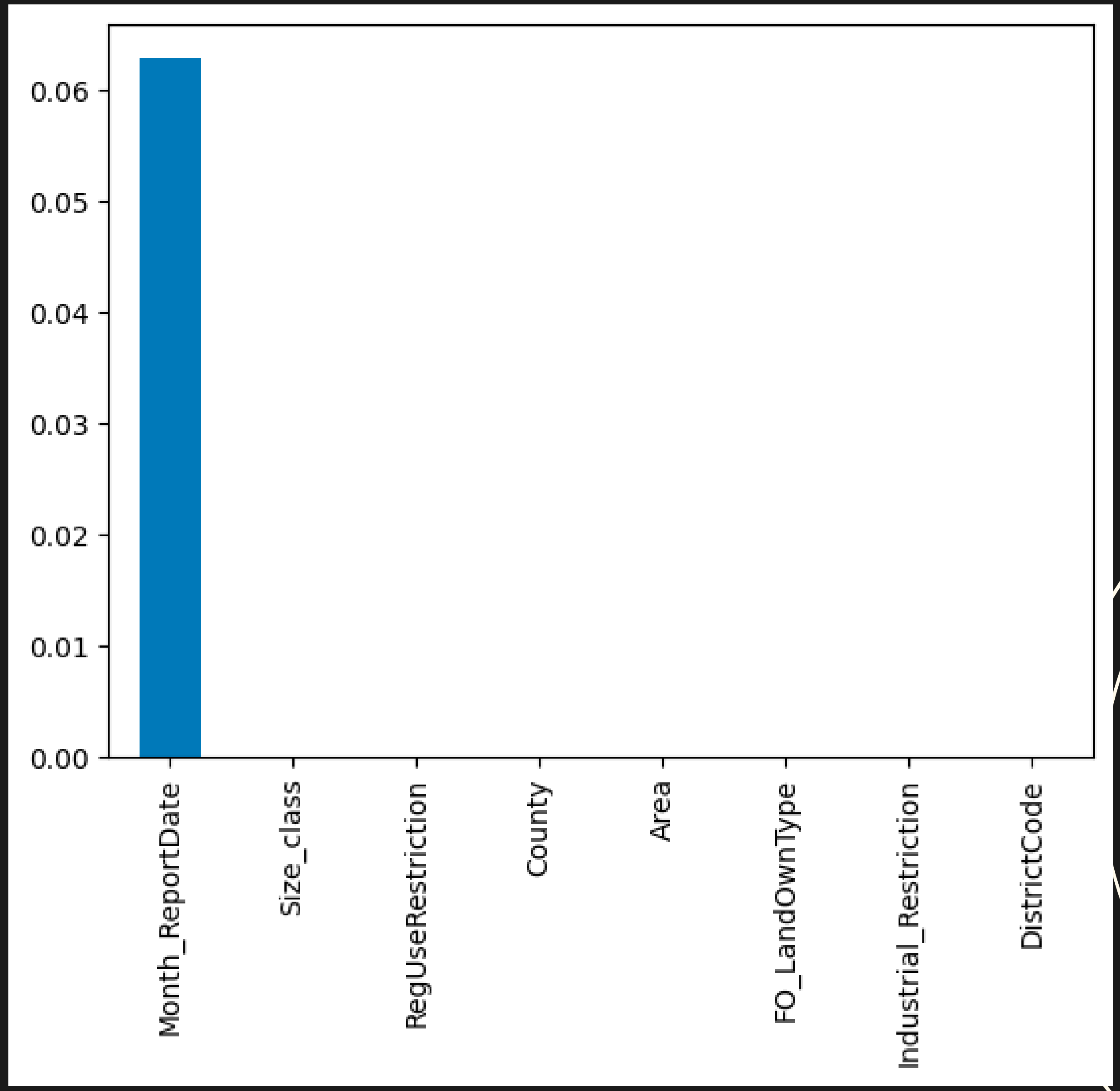
Train and test samples

test_size=0.3,
random_state=42

FEATURE SELECTION: CHI SQUARE

Useful when working with categorical or nominal data





Chi Scores

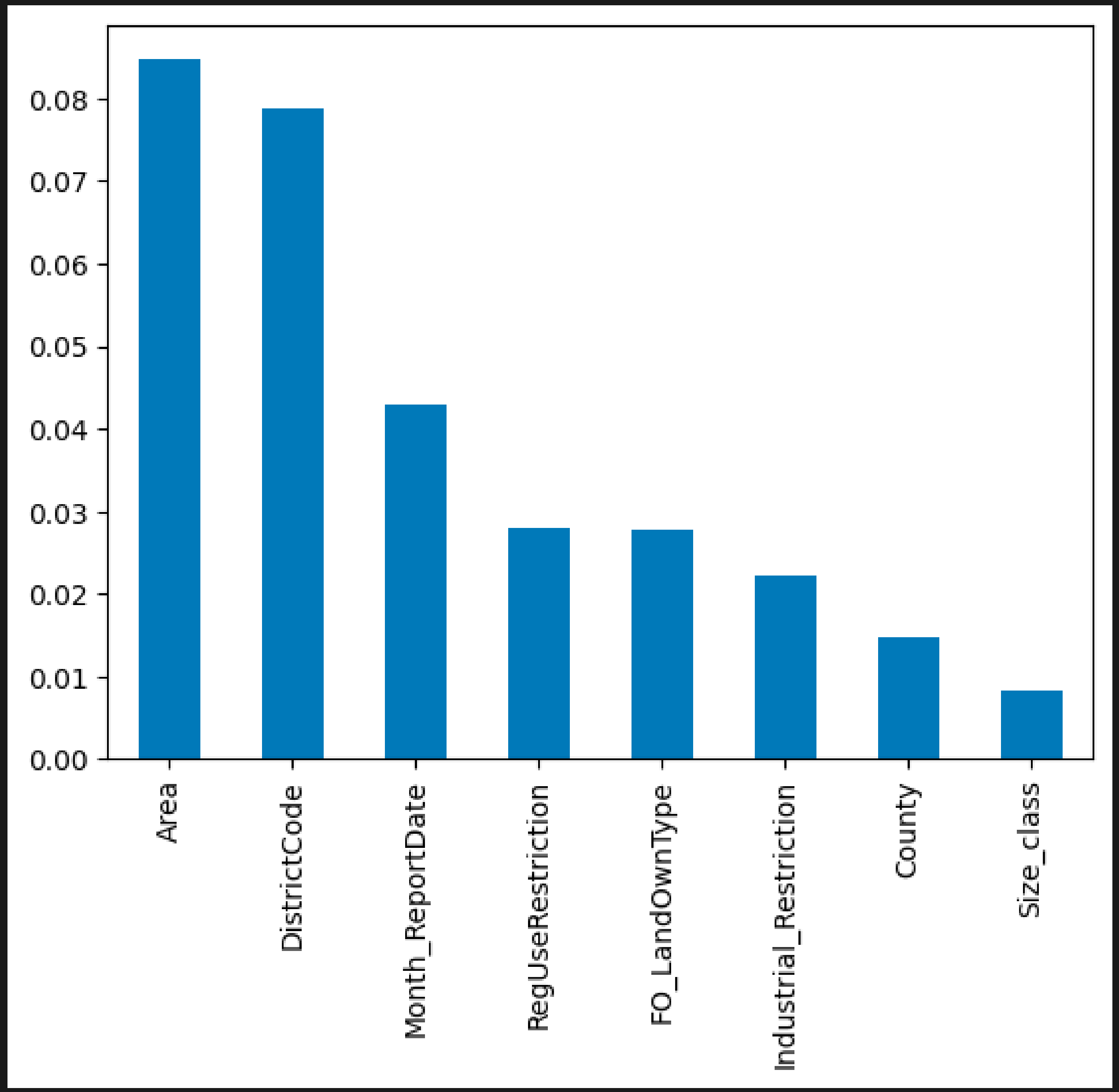
A higher value indicates a greater dissimilarity between the feature and target variable, suggesting a potentially significant association.

Chi p-values

Smaller p-value indicates stronger evidence against the null hypothesis, the feature is likely to be dependent on the target variable.

FEATURE SELECTION: MUTUAL INFORMATION

Measures the amount of information that one feature provides about another feature



Mutual information scores

A higher mutual information value indicates a stronger relationship or dependency between the feature and the target variable. It suggests that the feature contains useful information for predicting the target variable.

MODELS TRAININGS

Random Forest

- **Multiple decisions not based on a single feature's importance.**
- Focus on accuracy and robustness.
- It can handle a large number of input features and handle interactions between them effectively.
- Can handle imbalanced datasets well, which could be useful if the distribution of the fire causes is skewed.

Logistic regression

- **Relationship between chosen features and target based on probability.**
- Is a simple and interpretable model.
- Provides probability estimates for predictions, allowing us to set different decision thresholds based on our requirements.
- It can handle categorical features.

MODELS TRAININGS

Decision trees

- **Decides based on a set of features/attributes present in the data.**
- They can handle both categorical features without requiring extensive data preprocessing.
- They can handle interactions between features effectively.

K - neighbours

- **Based on similar fires that ocurred gives new classifications.**
- It does not assume any specific relationship between the features and the target.
- It can handle categorical features effectively.
- Captures complex decision boundaries and handles non-linear relationships.

EVALUATION AND SELECTION

Cross Validation

Provides a more robust and reliable estimate of a model's performance, helps in avoiding overfitting, supports hyperparameter tuning, facilitates fair model comparison,

Evaluation metrics

Accuracy
Precision
Recall
F1
Confusion matrix
ROC

V1

- Bad understanding of our dataset
- Wrong selection of features
- Wrong model selection.

V2

- Best understanding of features, but kept repetitive features.
- We didn't perform any data labelling for our categorical data.
- No feature selection method was implemented.
- Chose models without criteria.

V3

- Correct feature analysis **avoiding repetitive information.**
- Performed **feature labelling** and get dummies.
- Implemented 2 **feature selection methods: chose 3 features**.
- Better understanding of models and we chose **4 models with 5 evaluation metrics.**

V4

- Same as V3 but using just **2 features** trying to improve performance.

V5

- Same as V3 but n**ot using with get dummies** to increase performance.

Version	Details	Random forest	Logistic regression	Decision tree	k - neighbours
V3	FO_LandOwn Type, Industrial_Re striction, DistrictCode	Accuracy: 0.7956141958340969 Precision: 0.6847662141779789 Recall: 0.41978733240869165 F1 Score: 0.5204929779306392 Confusion Matrix: [[11209 836] [2510 1816]] AUC-ROC score: 0.6751904698573138	Accuracy: 0.8311037810762935 Precision: 0.6904122956818737 Recall: 0.6541840036985668 F1 Score: 0.6718100890207714 Confusion Matrix: [[10776 1269] [1496 2830]] AUC-ROC score: 0.7744145423225087	Accuracy: 0.7978132062793964 Precision: 0.5942486085343228 Recall: 0.7404068423485899 F1 Score: 0.6593248250308769 Confusion Matrix: [[9858 2187] [1123 3203]] AUC-ROC score: 0.7794188632664494	Accuracy: 0.8144279518661047 Precision: 0.6559322033898305 Recall: 0.6262135922330098 F1 Score: 0.640728476821192 Confusion Matrix: [[10624 1421] [1617 2709]] AUC-ROC score: 0.7541196645266335
V4	Industrial_Re striction, DistrictCode	Accuracy: 0.7614073666849918 Precision: 0.5699533644237175 Recall: 0.39551548774849743 F1 Score: 0.466975982532751 Confusion Matrix: [[10754 1291] [2615 1711]] AUC-ROC score: 0.6441670423383418	Accuracy: 0.763789628000733 Precision: 0.5533845080251221 Recall: 0.5499306518723994 F1 Score: 0.5516521739130434 Confusion Matrix: [[10125 1920] [1947 2379]] AUC-ROC score: 0.6952642051391885	Accuracy: 0.7625068719076415 Precision: 0.5514809590973202 Recall: 0.5423023578363384 F1 Score: 0.5468531468531468 Confusion Matrix: [[10137 1908] [1980 2346]] AUC-ROC score: 0.6919481901261392	Accuracy: 0.7559709241952233 Precision: 0.5445251546946462 Recall: 0.4678687008784096 F1 Score: 0.5032947905010569 Confusion Matrix: [[10352 1693] [2302 2024]] AUC-ROC score: 0.6636562267364237
V5	FO_LandOwn Type, Industrial_Re striction, DistrictCode No get dummies	Accuracy: 0.8213304013194063 Precision: 0.6596035543403964 Recall: 0.6692094313453537 F1 Score: 0.6643717728055079 Confusion Matrix: [[10551 1494] [1431 2895]] AUC-ROC score: 0.7725872810525024	Accuracy: 0.7560320078187038 Precision: 0.5854788877445932 Recall: 0.26282940360610263 F1 Score: 0.36279514996809187 Confusion Matrix: [[11240 805] [3189 1137]] AUC-ROC score: 0.5979983464688878	Accuracy: 0.7797935373526358 Precision: 0.5841306884480747 Recall: 0.5785945446139621 F1 Score: 0.5813494367669261 Confusion Matrix: [[10263 1782] [1823 2503]] AUC-ROC score: 0.715324669567255	Accuracy: 0.8101520982224666 Precision: 0.6481751824817519 Recall: 0.6158113730929264 F1 Score: 0.631578947368421 Confusion Matrix: [[10599 1446] [1662 2664]] AUC-ROC score: 0.7478807799462142

CHALLENGES

Evaluation metrics

Reading precision and recall for both final classifications since either lighting or human has more weight.

	REAL VALUES	
PREDICTED VALUES	TN	FP
	FN	TP
	10776	1269
	1496	2830

Human correcly labeled

Lightning correcly labeled

Precision	0.6904123	Precision	$P = \frac{TP}{TP + FP}$
Recall	0.654184	Recall	$R = \frac{TP}{TP + FN}$
	0.87809648		$P' = \frac{TN}{TN + FP}$
	0.89464508		$R' = \frac{TN}{TN + FN}$

Inverted, here we use TN as the base.