**Predicting Wildfire Damage in California:**

**Introduction**

This report presents a predictive model using machine learning techniques, specifically k-Nearest Neighbors (kNN) and Naïve Bayes, to estimate the probability of structural destruction (>50% damage) due to wildfires. The dataset used contains information on various property attributes, including year built, assessed value, construction materials, and geographic coordinates (latitude and longitude).

**Data Preparation**

Handling Missing Data:

* Homes built before 1900 (10,621 entries) were removed due to inconsistent records. Most pre-1900 records had values of zero, making it unclear which year they actually belonged to. However, the observed damage rate for these entries was lower than for post-1900 structures. Given that lower-damage data might lead to underestimation of fire risks, excluding them helps ensure a more cautious predictive approach.
* Missing values in year built were imputed using the median of post-1900 homes because the damage rate for missing values was similar to that of post-1900 structures.
* Assessed value had minor missing values (~6%) and was dropped to avoid imputation bias.
* The county feature was excluded from the model but used for visualization.
* The "unknown" category was retained in categorical variables like roof construction and exterior siding because its damage rate was similar to materials with damage rates in the lower end, indicating it carried useful information rather than being pure noise.

Features used:

* Features were selected using Mutual Information (MI), which measures how much information a feature contributes to predicting wildfire damage.
* The top two features with the highest MI scores—roof construction and exterior siding—were chosen, as they had strong predictive power for fire damage.
* These features had moderate missing values, making them a balanced choice between informativeness and data availability.
* The final feature set was selected to maximize predictive power while minimizing the impact of missing values, ensuring that key structural characteristics influencing fire risk were retained.
* Categorical variables were converted into numerical representations using one-hot encoding.
* The dataset was normalized to ensure consistency in scale, particularly for numerical variables.

**Model methodology and results**

* We compared two classification models to predict wildfire damage:
* Naïve Bayes Classifier – A probabilistic model assuming feature independence.
* k-Nearest Neighbors (kNN) – A distance-based model that classifies homes based on similar neighbors.
* The dataset was split into 70% training and 30% test data. Cross-validation was applied using a k-fold approach - systematically splitting the training dataset into multiple training and validation subsets. This process allowed every data point to be used for both training and validation at some point, reducing bias and increasing generalization. The 30% test set was held out for final evaluation.
* The optimal k-value (number of neighbors) was determined as 18 using cross-validation, ensuring the best balance between bias and variance.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **kNN** | **Naïve Bayes** | **Explanation** |
| Accuracy | 87% | 77% | Percentage of all predictions (damaged and non-damaged) that were correct. |
| Precision (Damage >50%) | 89% | 88% | When the model predicts a home is damaged, this is the percentage of those predictions that are correct. |
| Recall  (Damage >50%) | 89% | 69% | Out of all truly damaged homes, this is the percentage the model correctly identified as damaged. |
| F1-score | 89% | 77% | The harmonic mean of precision and recall, balancing both metrics for a more comprehensive evaluation. |

**kNN outperformed Naïve Bayes in terms of accuracy and recall, making it the preferred model.**

Confusion Matrix for kNN model - Insights

* False Negatives: It missed 1,597 truly damaged homes, leading to an 11% miss rate.
* False Positives: The model incorrectly flagged 1,581 non-damaged homes as damaged.
* While the false negatives are higher than false positives, kNN’s high recall for damaged homes makes it a reliable model. Fine-tuning could further minimize missed cases and improve prediction performance.

**Geospatial Visualization**

A scatter plot of latitude vs. longitude was created to visualize fire damage risk. (Figure 5) Counties with the highest destruction rates were Mono, Lake, and San Luis Obispo (See figures 3 and 4), highlighting regions requiring stronger mitigation efforts.

**Conclusion & Recommendations**

The following recommendations can help improve future risk assessment:

* Feature Enhancement: Incorporate humidity, vegetation data, historical fire occurrences, type of fire etc. to refine predictions. Poor data quality led to many variables being unused.
* Threshold Tuning: Adjusting the classification threshold could improve recall and minimize false negatives.
* High-risk counties should be prioritized for improved building materials and fire-resistant designs (See figures 1 and 2).

**Key Graphs:**

**Fire damage > 50% vs house construction materials used (Figures 1 and 2)**

Using fire resistant building materials clearly made a difference in reducing damage.

**A chart of a fire damage distribution

Description automatically generatedA graph of a fire damage distribution

Description automatically generated**

**Fire damage rate – County Statistics (Figures 3 and 4):**

**A screen shot of a graph

Description automatically generatedA graph of different colored bars

Description automatically generated with medium confidence**

**Figure 5 – Scatter plot of damage levels by latitude and longitude:**

**A red and green dots

Description automatically generated**