Casualty and Injury Prediction With Decision Tree Regressor

Ayça Durmuş
Middle East Technical University
Ankara, Turkey
ayca.durmus.ad@gmail.com

***Abstract*—Fires causes fatalities and monetary loses and using machine learning techniques to predict casualties can help with resource allocation and therefore reduce harm caused by fires. In this paper, using 2023 Fire Intervention Database from Izmir, Turkey, a casualty prediction model will be designed using decision tree regressor.**

Keywords—casualty prediction, injury prediction, fire

I. Introduction

Fires and related catastrophes cost lives and lead to financial loss. Predicting fires, or at least predicting casualties caused by a fire, can be a useful method for preventing the loss of human lives and allocating enough resources where it is needed.

Hu, Hu, and Hou developed a high performance model for predicting casualties in the case of emergencies using Global Terrorism Database [1]. They trained and evaluated a 2-step model that can predict casualties in the case of terrorist attacks. It first uses classification technics to predict if an event will result in a casualty or not. Later, it uses regression technics to predict the number of casualties. The study shows that 2-step prediction model performs better than the regression models when used alone. Back propagation (BP) neural network regression combined with Random Forest classification gives the best results among the combination of models tested in the paper.

Although the 2-step technique performs better than the traditional single step models, it is not suitable for this paper due to the sparseness of the database used, thus a single-step model will be used.

Schmidt, Gemmil, and Hoskins focus on structure fire incidents taking place in Oregon from January 2012 to August 2023 [2]. They used bagged decision tree classifiers with random forest algorithm to quantify the impact of several factors on the severity of injuries. The most important parameters were identified as the following: fire service response times, availability of working smoke or fire detectors, and the age of victims. Finally, a predictive Bayesian regularized neural network ensemble classifier was developed to model casualty severity. This model was projected spatially on the census block level.

Liu and Zhuang developed a framework that combine tree-based machine learning techniques with resampling methods to understand urban residential fire risks better and to enhance damage and casualty prediction [3]. To address the issue of database imbalance, various resampling techniques were tested. Combination of random undersampling and SMOTE lead to the best model performance and greatly increased the F1 score. To identify the key features influencing the model outputs, explainable Shapley additive explanations were used to understand the inner mechanisms of the model. The study improved the strategies used for combatting residential fires and offers tailored policies to increase fire safety and reduce risks by highlighting the heterogeneity in different regions, causes of fires, types of living spaces, etc.

II. Research Questions

A. *RQ1 - Statistical Research Question*

In this paper, two research questions regarding 2023 Fire Intervention Database will be answered. Structural fire causes the majority of the civilian fire-related deaths and injuries [2], thus, this paper will try to understand if urban fires are more deadly than rural fires, as urban areas have more structures due to population density. This question will be answered using Mann-Whitney U test to see if there is a statistically significant difference between casualties / deaths caused by urban and rural fires.

B. *RQ2 - Machine Learning Research Question*

Second research question that will be analysed in this paper is related to casualty prediction. Due to the sparseness of the dataset, injuries and casualties of firefighters and civilians will be combined. The decision tree regressor model will try to predict the overall number of casualties and injuries. Animal deaths won't be added to this count because the dataset is mostly concerned with farm animals, thus it might make the casualty / injury number biased towards rural fires.

This paper aims to document an experimental, iterative process aimed to answer RQ1 and RQ2. While trying to answer RQ2, linear regression was first tried and then discarded due to low R-squared value. Due to the spareness of the data, the problem at hand is hypothesized to be non-linear, thus decision tree regressor was tried next to potentially capture the non-linearity of the problem.

III. Preprocessing

A. *Data Inspection / Format Fixing*

Inspecting the descriptive statistics of the data shows that majority of the data entries have zero values and that the data is right-skewed. In this section, minor data fixes were performed. Id column was dropped and column names were simplified and converted to lower case. String values were converted to lower case. Address area (adres_bolgesi) is a binary categorical value, therefore it was converted to boolean type.

Since the database is highly unbalanced, injuries and death values were low overall, thus death / injury related columns (male, female, firefighter) were aggregated. Arrival time (varis_suresi) had some values that were incorrectly formatted, they were converted to correct format. The hour was deleted from date (tarih) because all the entries had zero as the hour.

B. Missing Data

The database is highly populated. 11 entries were missing in the arrival time column (varis_suresi) and 3 were missing in water used (kullanilan_su_miktari). These rows were dropped, as their percentage is low enough to be disregarded. Structure type (yapi_sekli) was missing for 9870 entries. After confirming that this column was only missing for fire types that are not related to structures, I filled these missing values with ‘not applicable’.

C. Temporal Feature Extraction

I converted date and teams departure time (tarih and ekiplerin_cikis_saati) to simplified categorical values. The model won’t be sensitive enough to predict the casualties based on temporal or seasonal information due to its imbalanced nature, thus these temporal values were simplified.

Date was converted to the season the fire happens in and departure time was converted to time of day (morning, afternoon, evening, night). These simplified categories were binned in a subjective manner and can be partitioned differently.

D. Checking The Distribution of Data

None of the numeric values (arrival time, amount of foam used, water used, dry chemical powder used, number of animals and people harmed) follow a normal distribution and with the exception of arrival time and water used, 75% of the values are zero. This means we are dealing with a highly unbalanced database. Due to this, I did not remove outliers as I assume it would eliminate any values that could be useful for casualty / injured prediction, as the target value is highly unbalanced.

Maximum value for humans harmed is 13, while it is 17500 for animals. Maximum value for foam used is 15000 kilograms. Maximum value for water used is 1000 metres cube. Maximum value for dry chemical power used is 250 kilograms. Maximum value for arrival time seemed to be 1445 minutes (24 hours), which seems to be an entry error, thus was deleted from the dataset. The maximum value for arrival time is 263 minutes.

The distribution of the categorical values revealed the following:

- Structure, garbage, and grass are the most popular fire types in the database.
- An overwhelming majority of the fires are caused by cigarettes / matches, short circuiting electrical boxes, and open fire.
- An overwhelming majority of the fires have been put out at the beginning.
- When a structure fire is involved, we see that the fires took place in reinforced concrete buildings (betonarme).
- The following districts had the most fires: Karabağlar, Konak, Bornova, Menemen, Buca.
- Majority of the fires took place in urban areas (kent_merkezi).
- The firefighters mostly left the station during afternoon and evening.
- Most popular seasons for the fires are summer and fall.
- None of the categorical columns are uniformly distributed.

E. Reducing Categorical Data

Fire type (yangin_turu) and cause of fire (yangin_sebebi) columns both had a high number of unique values (46 and 31 respectively). To simplify and improve the model, both of these columns’ values were separated into broad categories (9 for fire type and 6 for cause of fire).

For fire type, the values were separated into the following categories: natural areas, garbage and open fires, fires in living spaces , commercial and industrial fires, transportation fires, electrical energy caused fires, fires taking place in agriculture and animal husbandry, park and recreation spaces, and other.

For cause of fire, the values were separated into the following categories: unknown, accident and neglect, on purpose, natural causes, fire and spark, and heat and thermal causes.

Both of these categorizations are subjective and could be partitioned differently. Moreover, values in these categories are not mutually exclusive. E.g.: a fire can be both arson and caused with a fire source, thus these categorization is not perfect, but is done to the best of the ability of the author.

F. Binning Target Variable

The data needed to be split to avoid data leakage before applying one hot encoding and robust scaling. In order to ensure all 3 of the train / test / validation sets had equal number of non-zero target variable values, stratified binning was performed. Two bins were used, one for zero values and one for non-zero values. Dataset was split using stratified split based on these bins.

80% of the data was converted to training data, while both testing and validation datasets had 10% of the original dataset.

G. One Hot Encoding and Robust Scaling

Regression models works exclusively with numeric data, thus

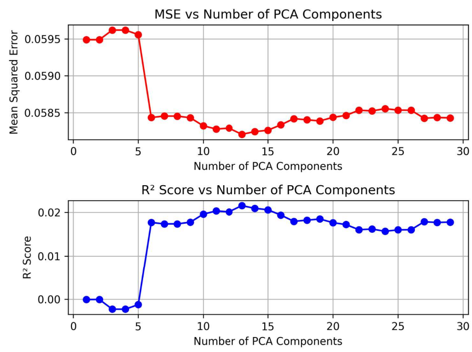

Figure 1: The MSE and F-squared Metrics According to Number of Principle Components

categorical values need to be transformed. One hot encoding was used for this task. One hot encoder was fit to the training data and all 3 datasets were transformed using this encoder. In order to get better model performance and to use PCA for dimension reduction, the numerical data need to be transformed. Because normalization and standardization are sensitive to outliers, robust scaler was used, as it takes outlier data into account. However, since I have switched to decision tree regressor later on, robust scaling was discarded.

H. PCA Dimensionality Reduction

Due to the one hot encoded categorical data, the dataset's dimension increased. In order to get a faster and better model, PCA was performed to lower the dimension size.

Different number of principle components (1 to 30) were tested using a linear regression model and evaluated with MSE (mean squared error) and R-squared metrics. The metrics show that best model performance was achieved with 13 principle components. However, both the MSE and R-squared metrics are not optimal even for the best number of principle component (0.0582 and 0.021, respectively). Due to these low metrics, I have switched to decision tree regressor for RQ2. Moreover, in order to interpret the results, I didn't use the PCA transformed datasets.

I. Log Transformation of Target Variable

In order to increase the model performance, I have transformed the target variable using numpy log transformation. I have used (x+1) natural logarithmic transformation as it is better for transforming skewed data and normalizing it.

IV. RQ1 – MANN WHITNEY U TEST

The human casualty / injury column of the data is not normally distributed. Therefore, to answer our statistical question we need a non-parametric test. RQ2 compares two groups that are not paired, thus Mann-Whitney U was chosen for the RQ2. Using the Shapiro test for normalcy, both the urban and rural casualty data was tested for normalcy. However, Shapiro test does not work accurately when the sample size exceeds 5000, thus I have randomly sampled 3000 samples from the both groups to check for the distributions. Both groups were confirmed to have non-normal distributions.

Significance level was chosen from papers that are related to fire statistics, casualty prediction, and perceived risk regarding residential fire injuries [4][5][6]. It was set to 5 percent.

After performing the Mann-Whitney U test, it was concluded that urban fires lead to more casualties.

V. HYPER PARAMETER TUNING

A. Model 1

To find the optimal hyperparameters for decision tree regressor model, I have used random search CV method and randomly searched 10.000 combinations of the parameters and the values seen in Table 1.

Table 1: Hyperparameters and values tested for the tree regressor

Hyperparameter	Values
Criterion	Squared error, Friedman MSE, absolute error, Poisson
Splitter	Best, random
Max Depth	5,10,15,20,15,30,35,40,50,60,70,80
Minimum Samples Per Split	2,4,8,16,32
Minimum Samples Per Leaf	1,2,4,8,16,32
Minimum Weight Fraction Per Leaf	0.05, 0.10, 0.15, 0.20, 0.25, 0.30
Maximum Features	2,4,8,16,32

Additionally, I assigned different weights to zero and non-zero values in the target value. Non-zero values were approximately 1% of the training set, thus I have assigned 1 for the zero values and 100 for the non-zero values. Combined with the minimum weight fraction per leaf hyperparameter, the goal is to give more importance to non-zero values.

The MSE of this model is 0.166 and it predicts either 0 or 1. It is an improvement from the linear regression model, however, as it will be discussed in the following section, it only creates a tree of depth 2.

Its best hyperparameters is the following:

Table 2: Best hyperparameters for model 1

Hyperparameter Type	Best Value
Criterion	Absolute Error
Splitter	Best
Max Depth	70
Minimum Samples Per Split	16
Minimum Samples Per Leaf	16
Minimum Weight Fraction Per Leaf	0.3

B. Model 2

In order to create a model with a higher depth, I decreased the weight given to non-zero target values by half, and added 0.005, 0.01, and 0.03 values to minimum weight fraction per leaf hyperparameter. The goal is to create a tree with a higher depth and better prediction capabilities by lowering the restriction created by the sample weights. However, this model led to a tree of depth 1.

Table 3: Best hyperparameters for model 2

Hyperparameter Type	Best Value
Criterion	Absolute Error
Splitter	Random
Max Depth	40
Minimum Samples Per Split	16
Minimum Samples Per Leaf	16
Minimum Weight Fraction Per Leaf	0.3

C. Model 3

For the third and final model, I didn't use any sample weights and didn't try to tune minimum weight fraction per leaf. Rest of the parameters were the same as the previous models. This model led to a tree of proper depth, however it could only predict zeros.

Table 4: Best hyperparameters for model 3

Hyperparameter Type	Best Value
Criterion	Absolute Error
Splitter	Best
Max Depth	60
Minimum Samples Per Split	8
Minimum Samples Per Leaf	4
Minimum Weight Fraction Per Leaf	-

Since model 1 is the only one that can produce non-zero predictions, I will use it for section 7. For the following section, all 3 models will be analysed for interpretability.

VI. INTERPRETABILITY

A. Model 1

Model 1 predicts injuries / casualties based on fire type. If the fire type is category 3, the model predicts casualty / injury as 1. Category 3 includes structure fires and living spaces. This finding is consistent with the literature, as structural fires lead to more deaths and injuries [2].

B. Model 2

As model 2 does not have any splitting points, it doesn't have any features given importance.

C. Model 3

Model 3 can only predict zeros, however it has the most complex tree out of all the models experimented here. Refer to Table 5 for the 10 highest feature importance scores.

Table 5: Feature importance scores of model 3

Feature Name	Score
yangin_turu_category 3	0.60952
kullanilan_su_miktari	0.59429
cikis_zamani_afternoon	0.23619
yangin_sonucu_başlangiçta söndürülen	0.15238
ilce_güzelbahçe	0.02952
yangin_sebebi_category 6	0.01333
ilce_karabağlar	0.01143
kullanilan_kopuk_miktari	0.00857
yangin_sebebi_category 2	0.00857
yapi_sekli_betonarme	0.00762

There are 2 features related to location in table 5. Location of fire is an important indicator for harm prediction [2][3], therefore these features are consistent with the literature. Moreover, living spaces as fire type has the highest score in model 3, similar to model 1.

Time of day the fire happens can impact the model prediction [2], thus cikis_zamani_afternoon (time of firefighters departure) is consistent with the literature.

Fire service response time is reported to be important for casualty severity [2], thus I was expecting to see similar results in my models also.

VII. Comparison with a Baseline

I have constructed a baseline model to compare the performance of model 1. The baseline model predicts zero, the dominant target variable, for all the data points in the test set. In order to compare the final predictions made by my model based on the test set, I have calculated the absolute error and squared error of all the predictions in both model 1 and the baseline model. Then, I have checked if these resulting datasets were normally distributed. I have used Shapiro test to check their normalcy. Since none of them were normally distributed, I used Wilcoxon signed-rank test to see if the errors for model 1 were less than the baseline model in a statistically significant amount. I have used Wilcoxon test because I am using the same test dataset to compare two models, thus I need a paired test that works with non-normally distributed data.

I have chosen to compare the means of absolute error and squared error because absolute error is less prone to outliers, and squared error penalizes bigger errors more harshly. I wanted to observe if there was a difference between performance metrics if I penalized outliers more, as the target variable is highly skewed and non-zero values can be considered as outliers.

My null hypothesis was that the difference between metric scores were zero majority of the time. My alternative hypothesis was that the difference between metric scores were less than zero majority of the time. This means my alternative hypothesis was claiming that model 1's absolute error and squared error values were lower than the baseline model.

The p-value results for absolute error and squared error comparisons were higher than the significance level (both scored 1, significance level is 5%), thus we can conclude that model 1 performs worse than or same as the baseline model. We can conclude the same when we check the mean of both absolute error and squared error metrics, as seen in Table 6.

Table 6: MAE and MSE scores of model 1 and baseline model

	MAE	MSE
Model 1	0.13	0.12
Baseline	0.02	0.06

VIII. Implications of the Study

The models I have experimented with did not perform better than a baseline model, thus the models' themselves can not be considered as contribution to the literature. However, through the analysis of feature importance (see section VI) and the result of the RQ1 (see section IV), we can conclude that the findings support the literature findings.

Section IV concludes that urban fires lead to more casualties and injuries, supporting the data that structure fires lead to more casualties and injuries [2]. Since urban areas have more structures than rural areas, it is possible to suggest that this finding from 2023 Fire Intervention Database from Izmir, Turkey support the overall claim that structural fires lead to more harm than other types of fires.

Section VI illustrates the parallels between the findings of the model experimented in this paper and the literature. Model 1 and model 3's first split points support the claim that structural fires are an important indication of civilian harm [2]. Moreover, model 3 identified other important decision points, such as location, time of day, and fire service response time, all consistent with the literature [2][3]. However, considering that model 3 predicts the same as the naive baseline model, these findings are not conclusive.

IX. Conclusion and Discussion

It was possible to answer RQ1. However, regressor models for casualty prediction based on the 2023 Fire Intervention Database (RQ2) could not perform better than a naive baseline. We had sufficient data to answer RQ1, but the target variable for RQ2 were highly skewed and non-zero values were low.

The main limitation for RQ2 was the distribution of the target variable. Future work can mitigate this issue by experimenting with over / under sampling methods [3]. Alternatively, instead of developing a regression model, we could develop a classification model that predicts if there will be casualties / injuries. This change in research question can potentially result in a better performing model, even with models with similar complexity as decision tree regressor.

Alternatively, we could attempt the same research question (RQ2) with more complex models [1][2]. Complex machine learning models could potentially identify patterns that weren't found by more simple models (linear regressor and decision tree regressor) used in this study.

References

[1] X. Hu, J. Hu, and M. Hou, "A two-step machine learning method for casualty prediction under emergencies," *Journal of Safety Science and Resilience*, vol. 3, no. 3, pp. 243–251, Sep. 2022. doi:10.1016/j.jnlssr.2022.03.001

[2] A. Schmidt, E. Gemmil, and R. Hoskins, "Machine learning based risk analysis and predictive modeling of structure fire related casualties," *Machine Learning with Applications*, vol. 20, Jun. 2025. doi:10.2139/ssrn.4757205

[3] Z. Liu and Y. Zhuang, "An investigation using resampling techniques and explainable machine learning to minimize fire losses in residential buildings," *Journal of Building Engineering*, vol. 95, Oct. 2024. doi:10.1016/j.jobe.2024.110080

[4] A. M. Hasofer and I. Thomas, “Analysis of fatalities and injuries in building fire statistics,” *Fire Safety Journal*, vol. 41, no. 1, pp. 2–14, Feb. 2006. doi:10.1016/j.firesaf.2005.07.006

[5] T.-Y. Huang and Y.-S. Lin, “Investigation of factors and construction of statistical models on predicting life casualties in building fires in New Taipei City,” *The Proceedings of 11th Asia-Oceania Symposium on Fire Science and Technology*, pp. 359–377, 2020. doi:10.1007/978-981-32-9139-3_27

[6] A. Mankell and F. Nilson, “A study of differences in the perceived risk of attaining a residential fire injury,” *Fire Technology*, vol. 59, no. 4, pp. 1789–1804, Apr. 2023. doi:10.1007/s10694-023-01410-x