

Climate Risk Analytics: California Wildfire Prediction Designed for Arbol”

Objective



The goal of this project is to analyze California wildfire data to:



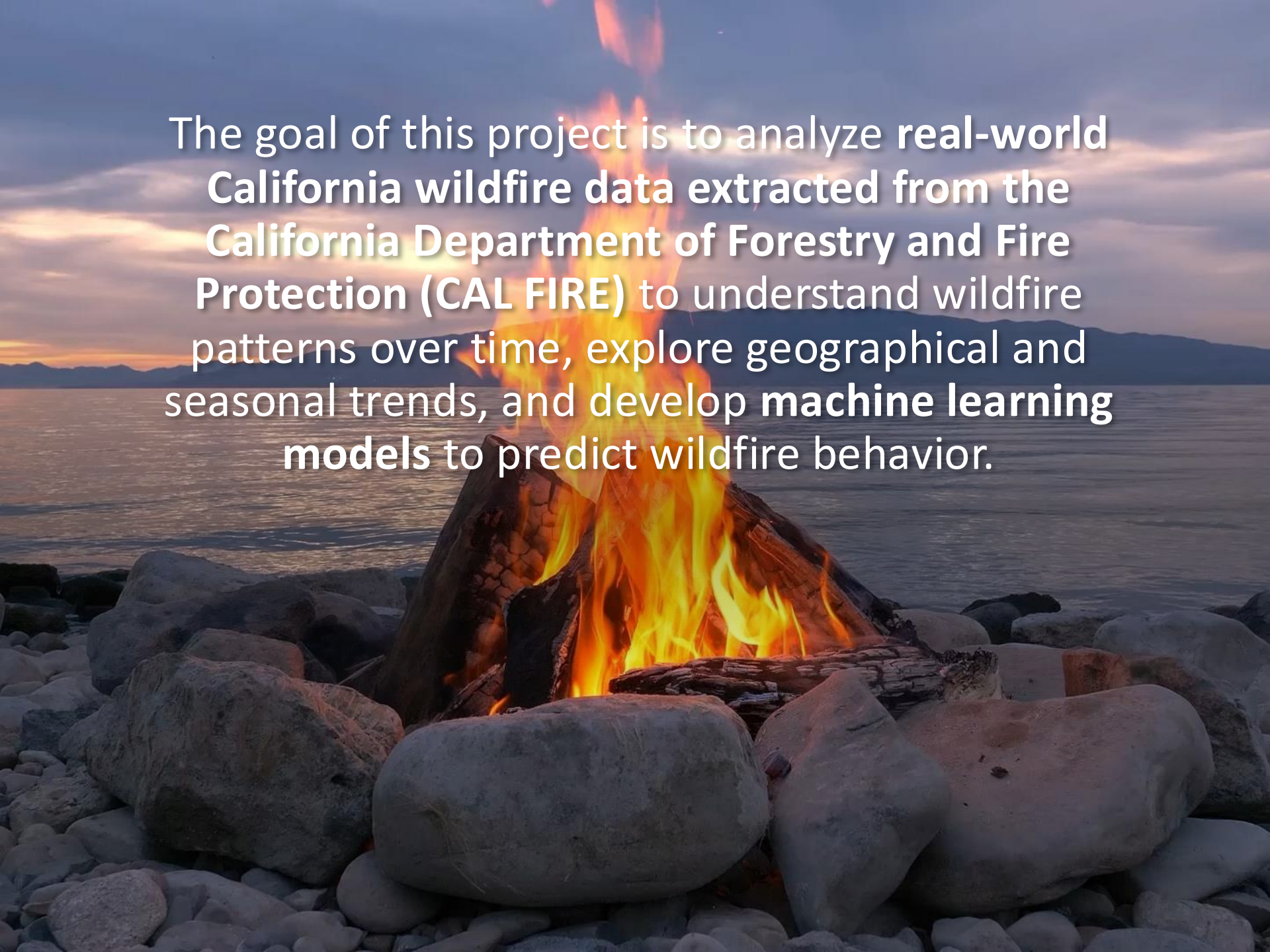
Understand wildfire patterns over time



Explore geographical and seasonal trends



Predict wildfire patterns using ML models

A campfire is burning brightly on a rocky beach. The fire is made of logs and is surrounded by large, smooth, grey rocks. In the background, the ocean is visible, and the sky is a mix of orange, yellow, and blue, suggesting a sunset or sunrise. The text is overlaid on the upper half of the image.

The goal of this project is to analyze **real-world California wildfire data** extracted from the **California Department of Forestry and Fire Protection (CAL FIRE)** to understand wildfire patterns over time, explore geographical and seasonal trends, and develop **machine learning models** to predict wildfire behavior.

Dataset Overview

The dataset includes wildfire incidents from 2009–2025 with:

Incident name

County & coordinates (latitude/longitude)

Acres burned

Year and month

Incident type



Data Cleaning Steps

We performed
the following
cleaning steps:

Removed
missing/null
values

Dropped
irrelevant
columns

Filtered dataset
by years of
interest

Ensured
correct data
types

Exploratory Data Analysis (EDA)

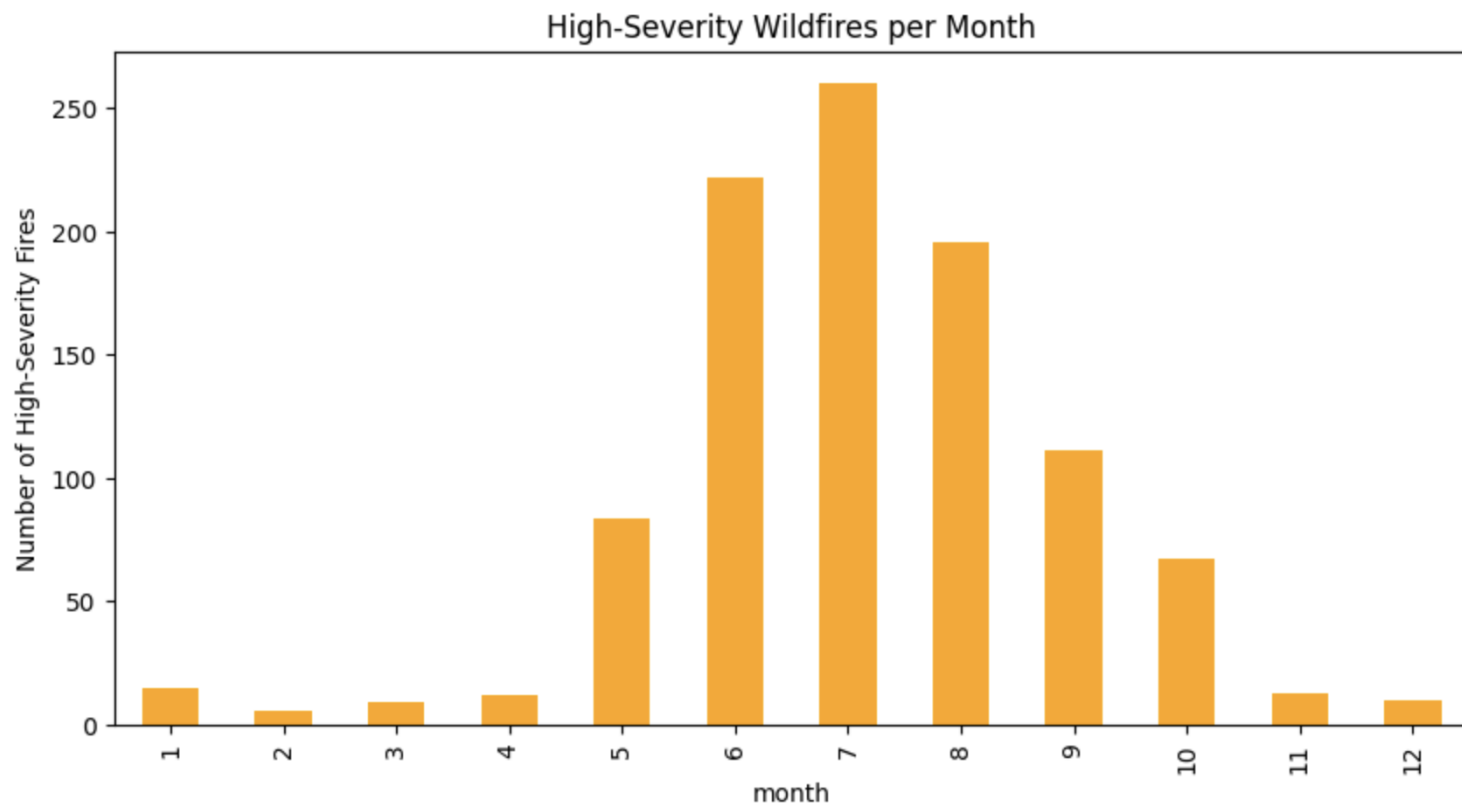
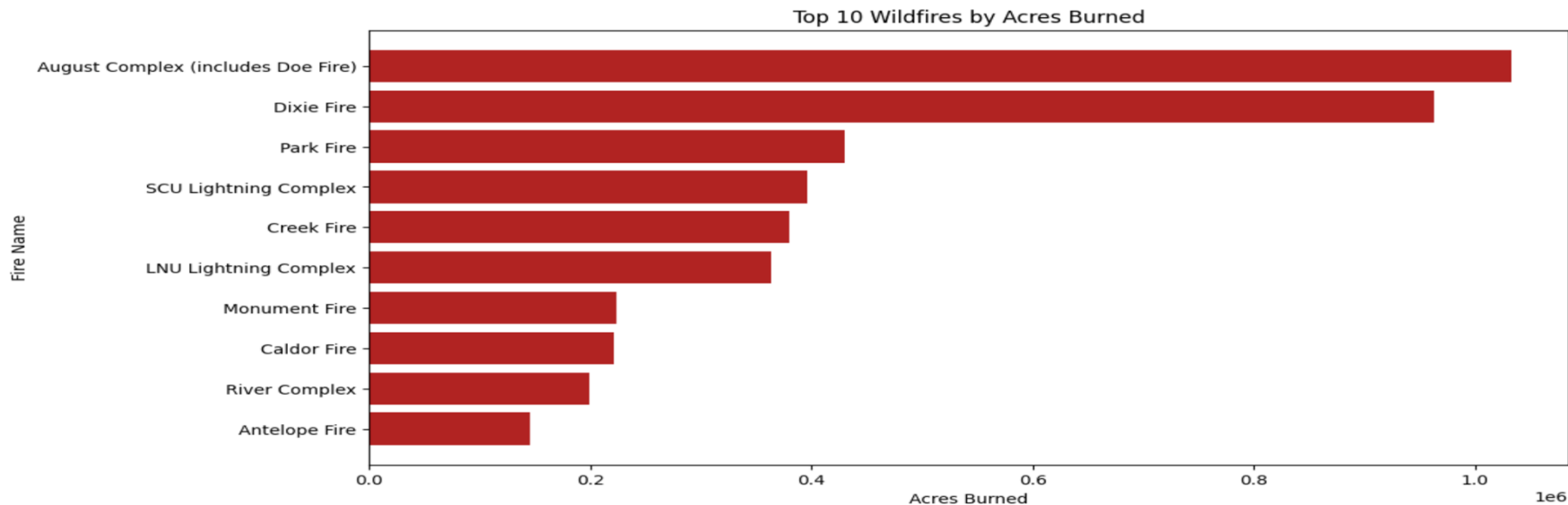
We explored the data using Python libraries:

Counted wildfire incidents per year

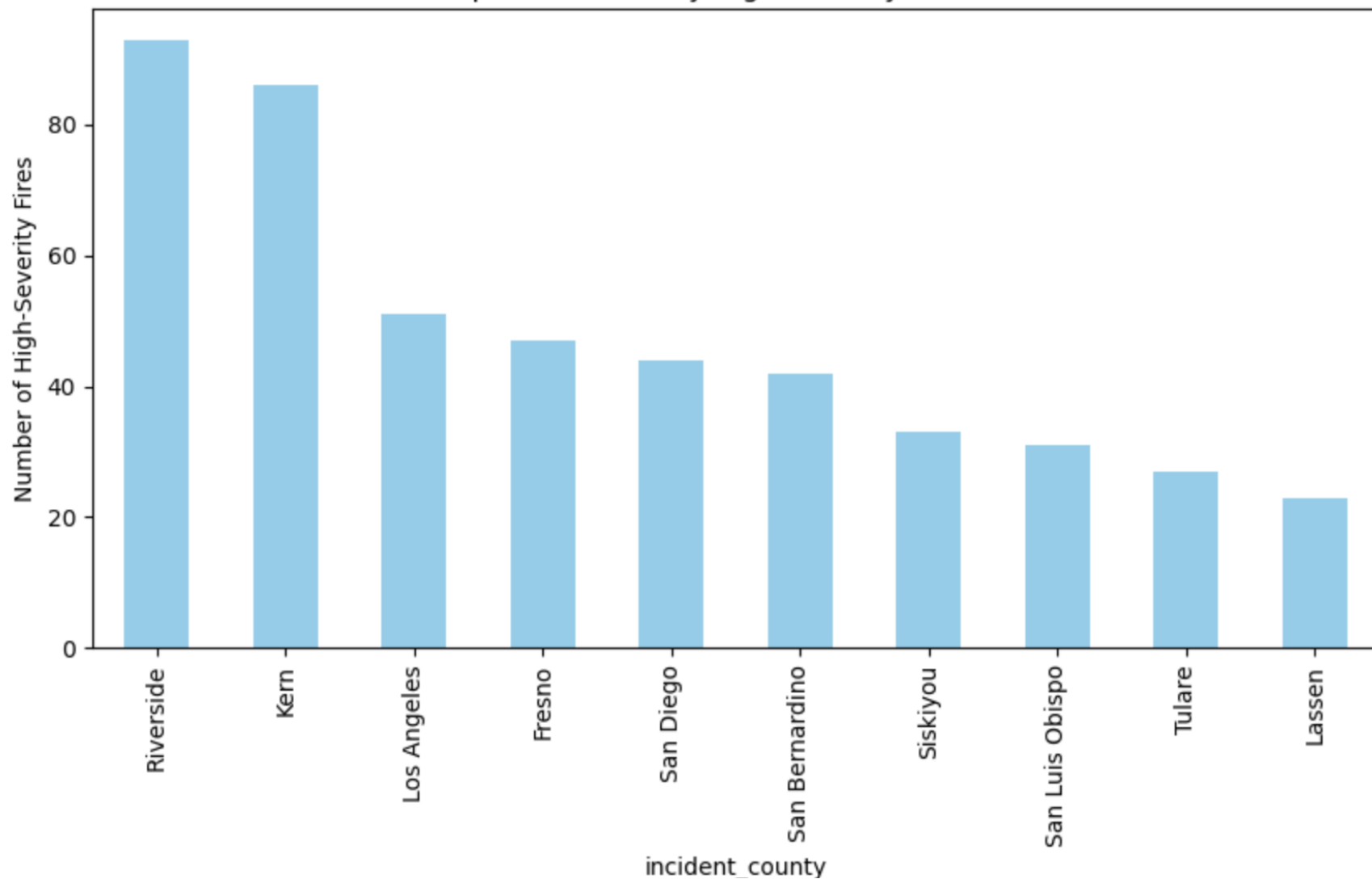
Identified most affected counties

Analyzed acres burned distribution

Checked most frequent fire names



Top 10 Counties by High-Severity Wildfires



Machine Learning Models

- We applied supervised ML models to predict wildfire patterns:

Predicting severity for 5 hypothetical fires:

Fire 1:

Location: (38.5°N, -120.5°W)

Time: Jul 2019

Predicted Severity:  HIGH SEVERITY

Confidence:

- High Severity: 76.5%
- Low/Medium Severity: 23.5%

Fire 2:

Location: (34.0°N, -118.2°W)

Time: Nov 2025

Predicted Severity:  LOW/MEDIUM SEVERITY

Confidence:

- High Severity: 32.3%
- Low/Medium Severity: 67.7%

Key Insights



FROM THE DATASET
ANALYSIS, WE
OBSERVED:



CERTAIN YEARS HAD
PEAK WILDFIRE
ACTIVITY



INCREASING
ACREAGE BURNED
OVER TIME



REGIONAL HOTSPOTS
MORE PRONE TO
WILDFIRES

Conclusion



The wildfire dataset provided valuable insights into patterns and impacts.



Data cleaning & EDA revealed trends over years



Further work needed with more diverse features



Supports predictive modeling for disaster management

Future Work

This model's accuracy of about 61% is expected given the difficulty of the problem and the limited feature set being used. It tries to predict a complex outcome wildfire severity using only location (latitude/longitude) and timing (year, month), without any direct information about critical drivers like weather, vegetation, fuel moisture, wind, topography, or firefighting response, which are known to strongly influence burn severity. Because many of these key variables are missing, a significant portion of the error is irreducible: even two fires in almost the same place and time can have very different severities due to unobserved factors. The dataset size (2,010 fires, perfectly balanced between high and low/medium severity) is reasonable, so the main limitation is not just “too little data” but that the available features carry only a modest amount of predictive signal. Despite this, the models achieve ROC-AUC values around 0.63–0.67, showing that they are meaningfully ranking fires by risk even if the raw accuracy at a fixed 50% threshold is moderate, which is still useful for parametric risk scoring and trigger design rather than exact yes/no classification

Model	Accuracy	ROC-AUC
Logistic Regression	0.614428	0.633821
SVM	0.609453	0.630133
Random Forest	0.604478	0.669736
Gradient Boosting	0.594527	0.628883



Best Model: Logistic Regression

Accuracy: 0.6144 (61.44%)

ROC-AUC: 0.6338