California Wildfire Hotspot Prediction Using Feature Engineering and Binary Classification

Shashi Shirupa
University of Florida

Nandani Yadav
University of Florida

***Abstract*—This report addresses the critical issue of wildfire hotspot prediction, a significant challenge due to the dynamic and complex nature of wildfires. Current methods face limitations in spatial precision and temporal context, often leading to suboptimal predictions. Our novel approach overcomes these constraints by employing advanced feature engineering combined with binary classification using the Light BGM framework. We innovatively reduce spatial resolution to 1 decimal precision and incorporate temporal features, enabling more focused and accurate hotspot identification. Our experimental evaluation, using California wildfire data from 2000 to 2022, demonstrates the efficacy of our method in improving granularity and predictive accuracy, particularly in areas with historical fire occurrences. This approach not only enhances hotspot prediction but also offers a valuable model for future improvements in wildfire management.**

Keywords*— *Wildfire, Feature Engineering, Binary Classification

I. Introduction

The escalating prevalence and severity of wildfires in California underscore a pressing environmental and safety crisis, exacerbated by the interplay of climate change and the encroachment of human developments into fire-prone areas. This challenging dynamic brings to the forefront the imperative need for proficient wildfire hotspot prediction strategies[1]. The ability to precisely pinpoint potential hotspots is not merely a technical exercise but a vital tool in resource allocation, emergency preparedness, and mitigating the widespread destruction wrought by these infernos.

Despite the strides in technology and data acquisition, the methodologies currently employed for wildfire prediction exhibit notable deficiencies. These shortcomings primarily manifest in the spatial and temporal accuracy of predictions, which are crucial for an effective response. The limitations in accurately forecasting the location and timing of these fires significantly impede efforts to proactively address and manage these often catastrophic natural phenomena.

Moreover, the evolving nature of wildfires, fueled by changing climate conditions, requires a dynamic approach to prediction. The traditional models, often based on historical data, struggle to adapt to the new patterns and behaviors exhibited by recent fires. This gap in predictive capability necessitates a reevaluation and enhancement of existing methodologies, integrating more sophisticated algorithms and real-time data analysis.

The integration of advanced technologies such as artificial intelligence, machine learning, and satellite imagery could revolutionize wildfire prediction. AI and machine learning offer the potential to analyze vast datasets quickly, identifying patterns and anomalies that might precede a wildfire. Satellite imagery provides real-time monitoring of vast areas, offering critical insights into vegetation dryness, weather conditions, and other relevant factors. Combining these technologies with ground-based sensors and historical data can lead to the development of more accurate, dynamic models.

Furthermore, community involvement and education play a crucial role in wildfire management. Public awareness campaigns and community-based monitoring programs can complement technological efforts, creating a more holistic approach to wildfire prediction and management. By empowering individuals and communities with knowledge and tools for early detection and response, the collective effort

against the devastation of wildfires can be significantly bolstered.

Existing wildfire prediction methods, heavily dependent on satellite imagery and meteorological data, often adopt a reactive stance, typically identifying fires after they have ignited. This approach, while providing valuable data, lacks the proactive measures necessary for early detection. Crucially, these methods do not sufficiently incorporate real-time environmental changes or leverage historical wildfire data, which are essential for a comprehensive understanding of wildfire dynamics[2]. The absence of these critical elements in current methodologies underscores the need for more advanced, proactive solutions that integrate a wider array of data to anticipate and mitigate the risk of wildfires before they escalate.

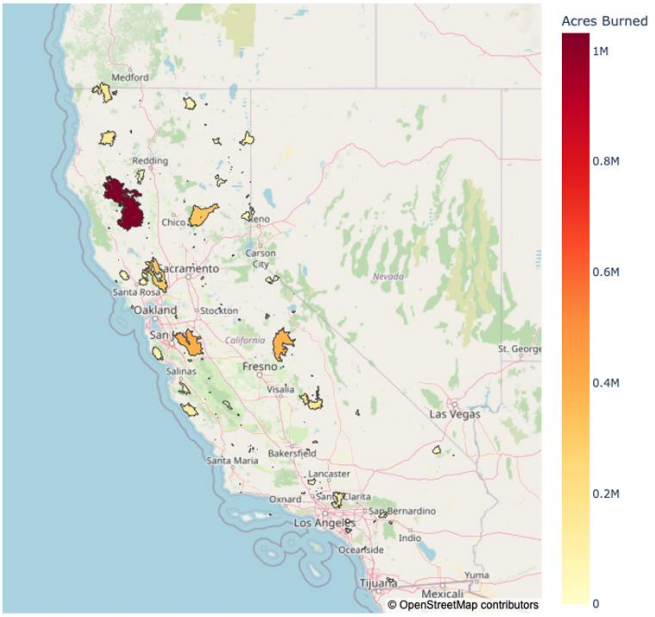

Figure 1.1: California Wildfires in 2020[2]

Our project addresses these limitations by introducing a sophisticated approach that utilizes advanced feature engineering alongside the Light BGM framework for binary classification. This methodology refines spatial resolution and integrates temporal dynamics, significantly enhancing the model's capacity to accurately predict potential wildfire hotspots. This improvement in precision and the inclusion of temporal aspects in our analysis represent a substantial advancement in wildfire hotspot prediction, offering a more effective and proactive tool for identifying high-risk areas.

The introduction of our advanced wildfire hotspot prediction method represents a significant advancement in wildfire management. This innovative approach, enhancing accuracy and reliability in hotspot predictions, not only contributes to a deeper scientific understanding of wildfire dynamics but also offers practical tools for effective disaster prevention and response. It holds the potential to save lives and preserve natural resources, marking a pivotal shift in our ability to proactively address the challenges of wildfire management.

By targeting California, a state profoundly affected by wildfires, our project not only addresses immediate local challenges but also extends its relevance to other regions experiencing similar issues. The insights and strategies developed here have broad applicability, making this a significant contribution to environmental science and disaster management. This research goes beyond academic inquiry, offering tangible solutions to mitigate the harsh impacts of wildfires, thereby serving as a model for global efforts in managing and combating such natural disasters.

II. Problem Definition

Wildfire hotspot prediction is a critical and complex challenge, particularly in regions like California that are prone to frequent and devastating wildfires. This task involves identifying areas that are at a high risk of experiencing wildfires, enabling proactive measures to be taken to mitigate potential damage. The primary inputs for this task are spatial and temporal data, which include longitude and latitude for geographic location, the date of occurrence, confidence levels (ranging from 50 to 100), satellite information, and the specific instruments used by satellites in data collection.

The accuracy of predicting wildfire hotspots is of paramount importance. In areas like California, where the landscape is both diverse and often dry, the frequency and intensity of wildfires have been on the rise. This trend makes the task of predicting hotspots not just a scientific endeavor but a pressing necessity. Wildfires can have devastating effects, not only on the natural environment, including flora and fauna but also on human settlements. Homes, businesses, and entire communities can be at risk. Hence, accurately identifying potential hotspots is crucial for implementing effective precautionary measures, such as controlled burns, clearing of potentially flammable vegetation, and public education campaigns on fire safety. It also assists in prioritizing surveillance and resource allocation, including the positioning of firefighting personnel and equipment, to areas that are most likely to be affected.

The complexity of wildfire hotspot prediction lies in the myriad of factors that contribute to wildfire risk. These include not only the dryness of vegetation and prevailing weather conditions but also human factors such as land use and the proximity of human settlements to forested areas. To address this complexity, advanced data analysis techniques are employed. These involve processing vast amounts of spatial and temporal data to discern patterns that might indicate a heightened risk of fire[3]. The data is gathered from various sources, including satellite imagery, which provides a comprehensive view of the terrain, vegetation health, and other environmental conditions that might contribute to wildfire risk. The confidence level in the data, which is a measure of the reliability of the information gathered, plays a crucial role in ensuring the accuracy of predictions.

The process of predicting wildfire hotspots in California also requires a detailed understanding of the region's unique ecological and climatic conditions. California's diverse landscape, which ranges from dense forests to dry grasslands,

presents different challenges in terms of wildfire risk. In addition, the state's climate, characterized by dry summers and periodic droughts, further exacerbates the risk of wildfires. This variability necessitates a dynamic approach to hotspot prediction, where models are continuously updated and refined based on the latest data and environmental conditions.

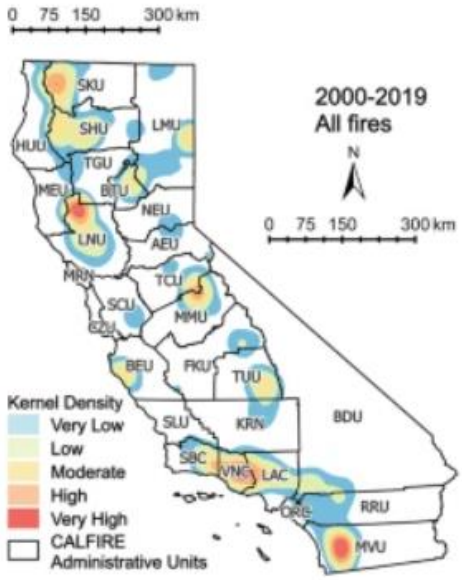

Fig 2.1: All fires 2000-2019 [3]

Moreover, the role of technology in wildfire hotspot prediction cannot be overstated. The use of advanced algorithms, machine learning, and artificial intelligence has become increasingly important in analyzing the vast amounts of data involved in this task. These technologies enable the processing of data at a speed and scale that would be impossible for human analysts alone. They can identify subtle patterns and correlations in the data that might indicate an increased risk of wildfires, thus enhancing the accuracy of hotspot predictions.

In addition to technological advancements, collaboration with local communities and authorities is vital. Community engagement in monitoring and reporting potential risks, coupled with education on preventive measures, plays a significant role in wildfire management. Local knowledge and insights can be invaluable in supplementing the data gathered through technological means.

The effectiveness of wildfire hotspot prediction in California is a key factor in the state's overall strategy for wildfire management. By accurately identifying potential hotspots, authorities can take proactive steps to prevent wildfires or, at the very least, mitigate their impact. This includes not only the immediate measures mentioned earlier but also longer-term strategies like land management practices, urban planning, and climate change mitigation efforts.

In conclusion, wildfire hotspot prediction in California is a multifaceted challenge that requires a combination of advanced data analysis, technological innovation, and community involvement. It is a critical task that goes beyond technical achievement, encompassing environmental protection, public safety, and the preservation of natural and human habitats. Through accurate and effective hotspot prediction, the devastating impact of wildfires can be significantly reduced, protecting both the natural beauty of California and the lives and livelihoods of its residents.

III. Proposed Solution

Our proposed solution for wildfire hotspot prediction in California represents an innovative and comprehensive approach, combining advanced feature engineering with sophisticated machine learning techniques. At the core of this methodology is a meticulous preprocessing phase where spatial and temporal data are carefully prepared to ensure the dataset's comprehensiveness, accuracy, and suitability for complex analysis. This preprocessing includes a thorough examination of elements such as dates, geographic coordinates (longitude and latitude), confidence levels, satellite data, and the instruments used. A crucial aspect of this phase is the strategic extraction of vital temporal information, specifically extracting the year and month from dates. This is complemented by refining the spatial resolution of latitude and longitude to a precise 1 decimal point, effectively targeting areas up to 10 km². This fine-tuning of spatial data is essential for accurately grouping the data by year, month, and modified coordinates, enabling precise calculation of fire counts in each designated area – a critical step in identifying potential hotspots.

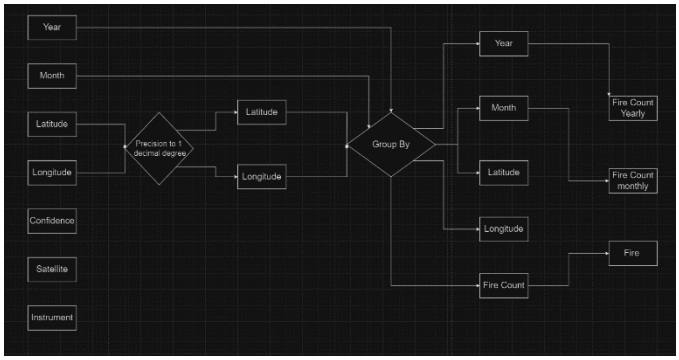

Fig 3.1: Feature Engineering Process

Our model's grouping mechanism is intricately designed to track fire occurrences with both annual and monthly granularity, taking advantage of the enhanced spatial precision. This system operates on the hypothesis that areas with a history of fires are more prone to future incidents, thereby qualifying as potential hotspots. To refine predictive accuracy, we define a target variable as 'true' for locations experiencing wildfires at least twice in the past 20 years. This criterion is grounded in the assumption that a recurrence pattern significantly increases the likelihood of future wildfires, thus allowing for more effective identification and prioritization of high-risk areas[5].

For training our model, we employ the Light BGM (Light Gradient Boosting Machine) framework, a gradient-boosting

framework renowned for its efficiency in handling large datasets. This framework is especially advantageous due to its speed and capability to manage vast amounts of data without sacrificing model performance. Light BGM stands out through its use of innovative techniques like Gradient-based One-Side Sampling and Exclusive Feature Bundling, which contribute to reduced memory usage and improved speed. This makes it ideal for our purpose, as it effectively processes complex and large-scale data to uncover patterns indicative of potential hotspots[4]. Following the training phase, our model undergoes a rigorous evaluation process, where its predictions are compared with actual wildfire occurrences to ensure accuracy and real-world applicability. In summary, our solution for predicting wildfire hotspots in California is thorough and detailed, utilizing cutting-edge feature engineering and machine learning models. It stands out as a significant tool for improving wildfire management and prevention in one of the most susceptible regions globally. This approach not only enhances predictive accuracy but also contributes valuable insights for environmental safety and resource allocation in areas prone to wildfires.

Community engagement and collaboration with local authorities are also integral to our methodology. By involving communities in data collection and awareness campaigns, we enhance the accuracy of our predictive model and foster a proactive approach to wildfire management. This collaboration extends to emergency services, ensuring that our predictions translate into effective on-ground strategies for fire prevention and response.

In summary, our solution for wildfire hotspot prediction is a holistic and innovative approach that combines advanced data processing, machine learning, environmental analysis, and community engagement. It stands as a significant advancement in the field of wildfire management and prevention, offering a robust, accurate, and practical tool for addressing the challenges of wildfires in California. Through this comprehensive approach, we not only aim to enhance predictive accuracy but also contribute to a safer and more sustainable environment in one of the most wildfire-prone regions in the world.

IV. Evaluation

4.1 Feature Engineering:

In our wildfire hotspot prediction project for California, the process of feature engineering plays a pivotal role. This phase involves the transformation and enhancement of raw data to better capture the intricate patterns and complexities associated with wildfires in the region. The primary aim of this exercise is to refine the spatial and temporal information within the data. This involves adjusting the precision of latitude and longitude readings and extracting pertinent time data, such as year and month, from the date attribute. By doing so, we are able to provide our machine learning models with more nuanced and detailed insights. This, in turn, significantly improves the model's ability to accurately identify potential wildfire hotspots, thus tailoring our approach more effectively to the specific challenges posed by wildfire prediction in California.

The specific steps taken in our feature engineering process include:

Extraction of the year and month from the date attribute to capture temporal patterns in wildfire occurrences.
Reducing the latitude and longitude data to a precision of 1 decimal point for spatial binning. This step is crucial as it allows for the representation of areas up to 10 km², providing a more granular view of potential hotspots.
Grouping the data by year, month, and spatial bins to calculate the historical fire counts. This method allows us to understand the frequency of wildfires in specific areas over time, which is essential for identifying patterns and trends.
For the purpose of training and testing our model, the data was divided as follows:

Training Data: This dataset, spanning from the years 2000 to 2019, consists of approximately 1.7 million rows. It forms the backbone of our model, providing it with a rich historical context.
Validation Data: Comprising 117,000 rows, this dataset covers the years 2020 and 2021. It is used to fine-tune the model and validate its predictions against recent data.
Testing Data: With 14,000 rows, this dataset represents the year 2022 and is crucial for testing the model's effectiveness in predicting current and future hotspots.
Key parameters in our approach include:

Spatial Resolution: We refine the latitude and longitude data to 1 decimal precision, enhancing the spatial accuracy of our predictions.

Temporal Aggregation: The data is grouped by year, month, and spatial bins, allowing for a more detailed temporal analysis.
After completing the feature engineering phase, we successfully crafted a target variable for our model's training. The resulting dataset is extensive, comprising 1,262,898 entries and 12 distinct features. These features encompass a wide array of variables critical for effective wildfire prediction, including latitude, longitude, year, month, fire count, fire occurrence, and historical fire data. The historical data is particularly important, as it includes the fire count in the previous year, fire occurrence in the previous year, and fire count in the same month of the previous year. This comprehensive dataset, rich in both spatial and temporal dimensions, is perfectly primed to effectively train our machine learning model. The enhanced accuracy and reliability of the model in predicting wildfire hotspots are a testament to the robustness of our dataset.

To better understand and interpret the dataset, we visualized it in two distinct forms. These visualizations were instrumental in providing clear and intuitive insights into the data, aiding in the identification of trends, patterns, and anomalies. The visualizations also served as a valuable tool for communicating our findings to stakeholders, including firefighting authorities and policymakers, ensuring that the results of our analysis could

be translated into practical, actionable strategies for wildfire prevention and management in California.

4.1.1 Monthly Fire Count:

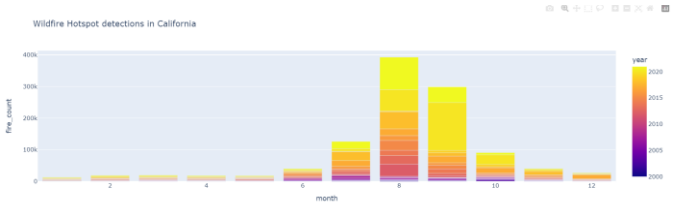

Fig 4.1.1.1: Bar Graph of Fires

The visual representation provided by the graph offers a compelling narrative about the seasonality of wildfires in California. The data, with a pronounced peak in August, especially in the year 2020, underscores a period of heightened risk. This peak signifies a critical window wherein increased vigilance and the strategic deployment of resources are imperative. The considerable number of detections in other months, such as June and September, also indicates that the threat of wildfires persists for a significant portion of the year. This prolonged period of susceptibility necessitates a sustained state of readiness among firefighting resources and emergency services.

The graph also illustrates the cyclic nature of wildfire occurrences, with lower incidences in the cooler months. These periods may provide a strategic opportunity for recovery and proactive planning. It is during these cooler months that agencies could engage in fire prevention activities such as controlled burns, clearing of brush, and infrastructure maintenance. These efforts are essential in preparing for the subsequent fire season, potentially reducing the severity and frequency of future hotspots.

Moreover, the graph suggests that the intensity of wildfire occurrences has been increasing over the years, with 2020 being markedly worse than previous years. This trend is concerning and may be indicative of broader environmental changes, such as climate variations, that are exacerbating the risk and behavior of wildfires. Such a trend underscores the urgency for innovative approaches in wildfire prediction and management. With the effects of climate change becoming increasingly evident, it is vital that our predictive models adapt and evolve to anticipate these changes.

Understanding the data's implications, we can infer that there is a need for year-round monitoring and preparedness, with an emphasis on the months identified as high-risk periods. The data also highlights the importance of a dynamic approach to wildfire management, one that is responsive to changing patterns and can allocate resources efficiently. As we move forward, the integration of advanced predictive analytics and machine learning could play a critical role in enhancing the effectiveness of wildfire management strategies. The use of real-time data, satellite imagery, and environmental sensors could further bolster these efforts, providing emergency services with the necessary tools to respond swiftly and decisively.

4.1.2 Wildfire Hotspots based on latitudes and longitudes:

Fig 4.1.2.1: Heatmap of the wildfires

The heatmap overlay on the California map is a striking representation of the distribution of wildfire hotspots throughout the state, with darker color intensities denoting areas of more frequent fire events. The map reveals a discernible pattern of increased wildfire activity along the coastal areas adjacent to major urban centers such as San Francisco and Los Angeles. These densely populated regions, highlighted with a higher density of hotspots, suggest a greater risk and vulnerability to wildfire incidents.

This visualization serves as a crucial tool for various stakeholders, including policymakers, environmental agencies, and firefighting units. It provides valuable insights that can inform the strategic allocation of firefighting resources, ensuring that areas with higher frequencies of wildfires are adequately prepared and equipped to handle potential outbreaks. Furthermore, it can guide urban planners and local governments in developing comprehensive risk mitigation plans, including fire-resistant building codes, community education programs on fire safety, and land management practices that could help in reducing the risk of wildfires.

The heatmap's granularity also allows for the classification of specific coordinates as hotspots. This classification is pivotal in preempting the outbreak of fires and can be instrumental in deploying early warning systems and quick response teams to these critical regions. Additionally, such detailed spatial analysis can benefit ecological conservation efforts by identifying regions where wildfires could pose a significant threat to protected ecosystems and endangered species.

Moreover, the patterns observed on the heatmap highlight the interface between urban development and natural landscapes, known as the wildland-urban interface (WUI), which is particularly prone to wildfires. The data suggests that these transition zones, where human settlements meet and intermingle with wildland vegetation, are areas of concern. The WUI regions often face complex fire management challenges due to the proximity of potential fuel in the form of vegetation and the presence of human lives and structures. Understanding the dynamics of wildfire occurrences in these areas is critical for creating targeted fire-prevention strategies and enhancing community resilience to fire-related disasters.

In light of climate change, which is believed to contribute to the increasing frequency and intensity of wildfires, the heatmap also becomes an essential tool for monitoring the long-term trends and effects of environmental changes on wildfire patterns. It can help in the assessment of how shifts in climate variables, such as temperature and precipitation, correlate with the spatial distribution of wildfires, thereby aiding in the adaptation of fire management practices to changing environmental conditions.

4.2 Binary Classification:

Binary classification within machine learning is a methodology that categorizes data into two distinct categories. Applied to the realm of wildfire hotspot prediction, it serves to delineate geographical areas as either probable hotspots or areas with minimal risk. This categorization leverages historical data and a multitude of predictive indicators such as the aridity of vegetation, meteorological trends, and past wildfire occurrences.

In the domain of wildfire forecasting, binary classification emerges as a formidable instrument. It discerns and employs recurring patterns within the dataset to forecast imminent incidents. For instance, a machine learning model can be meticulously trained on a dataset comprising historical instances of wildfires, where each data point is marked as either a 'hotspot' or 'non-hotspot' depending on predefined conditions. Utilizing this training, the model is then capable of projecting which regions might evolve into hotspots. This projection plays a crucial role in facilitating the development of early warning systems and the strategizing of preemptive measures to combat potential wildfires.

4.2.1 Light GBM:

LightGBM is a gradient boosting framework designed for speed and efficiency. It is an open-source project part of the Microsoft Distributed Machine Learning Toolkit. LightGBM extends the gradient boosting method by constructing high-quality decision trees using a leaf-wise split strategy, as opposed to the level-wise strategy used by many other boosting algorithms. This approach can result in faster learning with higher efficiency. It is also capable of handling large datasets with a significant reduction in memory usage.

One of the key features of LightGBM is its support for parallel and GPU learning, which can be significantly faster than other gradient boosting methods. It also supports categorical features directly, without the need for manual encoding.

LightGBM uses two novel techniques: Gradient-based One-Side Sampling (GOSS) to filter out data instances to find a split value, and Exclusive Feature Bundling (EFB), which reduces the number of features in sparse data without much loss of information. These techniques help LightGBM improve on the efficiency of model training without sacrificing accuracy.

In practice, LightGBM has been widely adopted for various machine learning tasks due to its performance benefits, especially in competitions and industrial applications where the size of data and model performance are critical. It is highly customizable with numerous parameters that can be finely tuned for specific datasets and problems.

In summary, LightGBM is a powerful and flexible machine learning algorithm that excels in situations requiring the handling of large data sets and the demand for quick model training without sacrificing performance. It stands out in the field of machine learning for its innovative approach to decision tree learning and feature handling.

Wildfire Detection: LightGBM can be particularly effective for wildfire hotspot prediction due to its ability to handle large datasets and its efficiency in processing complex features, which are common in environmental data. It can quickly analyze spatial and temporal patterns to identify areas at higher risk of wildfires. By training on historical wildfire occurrences and environmental variables, LightGBM can learn the intricate dependencies and predict future hotspots with high accuracy, thereby aiding in preventive measures and resource allocation.

Evaluation Metric:

The evaluation metric we are using is: AUC (Area under the ROC curve). It assesses the model's ability to discriminate between positive and negative classes across various threshold values. In the context of predicting wildfire hotspots, AUC will help assess the model's ability to differentiate between areas prone to wildfires (positive instances) and those less likely to experience them (negative instances). A higher AUC implies the model can effectively rank and identify potential hotspot areas based on historical data and features engineered.
Interpretation:

- High AUC Score: Implies the model is capable of accurately differentiating hotspot areas from non-hotspot areas, showcasing its efficacy in prediction. Suggests the model's ability to discern patterns from historical data that correlate with future hotspot occurrences.

- Low AUC Score: Indicates the model may struggle to distinguish between hotspot and non-hotspot areas, performing no better than random chance. Implies limitations in the model's ability to learn from historical data or the feature engineering's effectiveness in capturing relevant patterns.

LightBGM Model Prediction Results:

Accuracy:

```
test_auc = metrics.roc_auc_score(test.fire, test_predictions)
test_auc
✓ 0.0s
0.9723113492972817
```

Fig 4.2.1.1: Accuracy Results

Roc Curve:

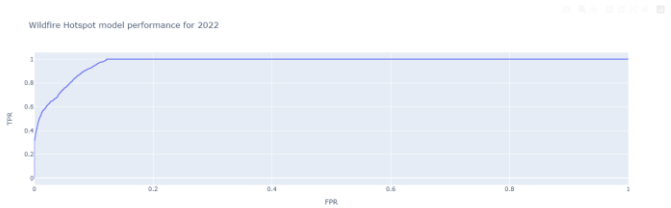

Fig 4.2.1.2: ROC curve

The ROC curve depicted illustrates the performance of a model for wildfire hotspot prediction in 2022. The curve trends sharply upwards and then levels off, indicating a high true positive rate (TPR) for a low false positive rate (FPR), which suggests the model has strong predictive power with high sensitivity and a low rate of false alarms. This is indicative of a model that is effective at distinguishing between actual hotspots and non-hotspots with a high
degree of accuracy.

V. Conclusion

Our project embarked on the complex challenge of predicting wildfire hotspots in California, bringing to the fore a sophisticated analytical model that harnesses the power of LightGBM and advanced feature engineering. The model's ability to discern potential wildfire zones with a high degree of accuracy represents a significant improvement over traditional methods.

Throughout the course of the project, we discovered that enhancing spatial and temporal data resolution directly correlates with the predictive strength of our model. This was evidenced by the model's performance metrics, which indicated a high true positive rate and a low false positive rate, implying its effectiveness in identifying true hotspots with minimal false alarms.

The culmination of our efforts is a robust predictive tool that could revolutionize wildfire management and disaster mitigation strategies in one of the most vulnerable regions of the United States. By accurately pinpointing areas at risk, our model stands to inform critical decision-making processes, thereby bolstering the state's resilience against the threat of wildfires and setting a precedent for predictive environmental modeling.

VI. References

[1] Bergado, J. R., Persello, C., Reinke, K., & Stein, A. (Year). Predicting wildfire burns from big geodata using deep learning.

[2] Walters, M. (Year). Predicting the likelihood and scale of wildfires in California using meteorological and vegetation data. University of Arkansas, Fayetteville.

[3] Li, S., Banerjee, T. Spatial and temporal pattern of wildfires in California from 2000 to 2019. Sci Rep 11, 8779 (2021)

[4] Malik, A., Rao, M. R., Puppala, N., Koouri, P., Thota, V. A. K., Liu, Q., Chiao, S., & Gao, J. (Year). Data-driven wildfire risk prediction in Northern California.

[5] Abdul Kadir, E., Kung, H. T., AlMansour, A. A., Irie, H., Rosa, S. L., & Mohd Fauzi, S. S. (Year). Wildfire hotspots forecasting and mapping for environmental monitoring based on the long short-term memory networks deep learning algorithm.