PREDICTIVE LEAD SCORING ENGINE USING MACHINE LEARNING

Preetha S[1], Nalini M K[2], Ibrahim Javeed Khan[3], Nanditha B Satish[4], Gagan V S[5]

Dept. of ISE, B.M.S. College of Engineering, VTU, Bengaluru, India

{preetha.ise,nalini.ise,ibrahimjaveed.is20,nandithab.is20,gaganv.is20}@bmsce.ac.in

ABSTRACT

Lead scoring is a process of assigning scores to prospect based on their behavior and interaction with the product or service. It is important for several reasons in the context of sales and marketing. Lead scoring enables businesses to prioritize and focus their efforts on leads that are most likely to convert into customers. By assigning a score or ranking to each lead based on their characteristics, behavior and engagement with the company's sales and marketing, teams can allocate their time and resources more effectively. Proposed system focuses on design and development of a web application for lead scoring. It discusses challenges of manual lead scoring, benefits of automation, and technical considerations involved in creating the web application. Study explores the required features, data security, and integration with existing systems. It analyzes technologies and algorithms used and outlines the testing and validation methodology. Findings highlight the importance of automated lead scoring in improving efficiency and conversion rates for sales and marketing teams. Research contributes to the field by providing insights into design and implementation of a web application for lead scoring along with aiding businesses in optimizing their lead management practices.

Keywords **—Lead Score, Prediction, Web Application, Machine Learning, AWS**

I. INTRODUCTION

In the fast-paced and competitive business landscape, effective lead management is crucial for organizations to generate and convert leads into customers. Lead scoring process evaluating quality and potential of leads, has emerged as a vital component of sales and marketing strategies. By systematically assigning scores or rankings to leads based on their characteristics, behavior, and engagement, businesses can prioritize their efforts, allocate resources efficiently, and maximize conversion rates. Traditional manual lead scoring methods have inherent limitations, including subjectivity, time-intensive processes, and potential inconsistencies. To overcome these challenges and improve the accuracy and efficiency of lead scoring, integration of technology, specifically web applications has become increasingly popular. Web applications provide automation capabilities that streamline lead scoring workflows, enhance data processing and analysis, and enable real-time insights for sales and marketing teams. Study focuses on the design and development of a web application for lead scoring, with the aim of addressing the challenges faced by businesses in manual lead scoring and exploring the benefits and technical considerations involved in creating an automated solution. An examination of necessary features and functionalities of a lead scoring web application, emphasizing the importance of data security and integration with existing systems, such as Customer Relationship Management (CRM) platforms.

Furthermore, study analyzes technologies and algorithms employed in the development of web app considering factors such as programming languages, frameworks, and data storage solutions. By providing insights into the design and implementation of a lead scoring web application, this study aims towards contribution to the existing body of knowledge in the field of lead management practices. The findings of this study have practical implications for businesses, as they highlight the significance of automated lead scoring in optimizing resource allocation, improving sales and marketing efficiency and ultimately driving higher conversion rates. Finally, web application ensures to make an end-to-end Software as a Service (SaaS) or Customer Relationship Management (CRM) tool. Currently study emphases in making a micro service which can be integrated later on.

II. RELATED WORK

An ML algorithm known as XGBoost was proposed in [1]. XGBoost is a powerful and scalable machine learning algorithm that belongs to the family of gradient boosting frameworks. It combines the strengths of decision trees and gradient boosting to deliver highly accurate predictions and handle large-scale datasets efficiently. Algorithm employed a novel regularization technique called "Gradient-based one-side Sampling" to enhance model generalization and prevent overfitting. XGBoost also utilizes parallel processing and distributed computing to optimize performance on systems with multiple cores. System gained popularity in various domains due to its superior predictive performance, flexibility, and speed. It has turned to a go-to choice for many data scientists and machine learning practitioners.

Authors in [2] discussed modeling lead scoring in digital marketing processes. Approach is valuable since it optimized and enhanced marketing efforts. By utilizing statistical and ma- chine learning techniques, businesses can develop predictive models that assign scores to leads based on their likelihood of converting into customers. This enables more targeted and personalized marketing strategies, allowing for efficient resource allocation and higher conversion rates. The models can leverage a variety of data sources, including demographics, browsing behavior and engagement metrics to generate ac- curate predictions. By incorporating lead scoring models into digital marketing processes, businesses can streamline lead qualification, improve lead nurturing strategies and ultimately drive better results in their marketing campaigns. In [3] Benhaddou et al. used Bayesian network elicitation techniques in customer relationship management (CRM) for building lead scoring models is a valuable approach that leverages small data. Small data refers to datasets with limited samples but rich information. By employing Bayesian networks, businesses can capture the complex relationships between various customer attributes and behaviors, enabling the development of accurate lead scoring models. These models take into account both quantitative and qualitative data, providing a holistic view of customer interactions. The application of Bayesian network elicitation techniques in CRM enables businesses to make in- formed decisions about lead prioritization, resource allocation, and personalized marketing strategies, leading to improved customer acquisition and retention. In [4] Slakey et al. proposed Bayesian models to overcome challenges posed by categorical variables for predictive modeling. By encoding categorical variables using conjugate Bayesian models, lead scoring engine can effectively handle categorical features and capture underlying relationships within the data. This method improved accuracy and performance of lead scoring engine, enabling evaluation and prioritizing leads more efficiently. Thereby leading to better conversion rates and enhanced decision-making in their lead scoring process.

Method in [5] pitches an AI-assisted lead scoring. It is a powerful approach that leverages artificial intelligence algorithms and techniques to optimize the process of evaluating and ranking potential customers. By analyzing a wide range of data sources including demographics, behavioral patterns, and engagement metrics, AI algorithms can automatically assign scores to leads based on their likelihood of conversion. This enables businesses to prioritize high-potential leads, personalize marketing strategies, and allocate resources more efficiently. AI-assisted lead scoring improves the accuracy and efficiency of lead qualification, enabling businesses to focus on leads with the highest probability of conversion and ultimately increase sales and revenue. A method used in [6] is based on "Predictive Lead Scoring". It focuses on addressing the challenge of imbalanced data in lead conversion prediction. The proposed method utilizes predictive lead scoring techniques to handle data imbalance and improve accuracy of lead conversion predictions. By considering various factors and attributes the method aims to create a model that can effectively evaluate and rank potential customers based on their likelihood of conversion. This approach contributes to optimizing lead qualification and enabling businesses to prioritize resources effectively; hence leading to improved conversion rates and more successful marketing strategies.

A versatile lead scoring model designed specifically for business-to-consumer (B2C) markets was introduced in [7]. Proposed model aims to evaluate and rank potential customers in an online setting based on their likelihood of conversion. By considering factors such as demographics, online behavior, and engagement history, the model enables businesses to effectively prioritize leads for targeted marketing strategies. Its generic nature allows for adaptability across various B2C industries and can be tailored to meet specific business requirements. Implementing this model facilitates efficient resource allocation, personalized customer interactions, and improved conversion rates in the B2C market. Authors of [8] focused on development of a lead scoring model tailored specifically for digital marketing firms operating in education sector in India. Model targeted to assist these firms in effectively evaluating and prioritizing leads based on their potential for conversion. By considering the unique dynamics and characteristics of the education vertical in India such as specific demographics and consumer behavior, the model can provide valuable insights for lead qualification. Implementing this customized lead scoring model enables digital marketing firms to optimize their strategies, allocate resources efficiently, and achieve better conversion rates in the education sector in India.

Jadli, Aissam et al. [9] explored development of an intelligent lead scoring system leveraging machine learning techniques. The objective was to create a system that can automatically analyze score leads based on various attributes and behaviors. The system utilized machine learning algorithms to uncover patterns and relationships in data enabling accurate predictions of lead conversion likelihood. This smart lead scoring system enhances lead qualification processes allowing businesses to prioritize high-potential leads, personalize marketing strategies and allocate resources more effectively. As a result, conversion rates were improved, sales and marketing efforts was optimized and better results in lead generation and customer acquisition was observed. A novel approach that combines feature selection techniques with the CatBoost algorithm for predicting aboveground biomass was proposed in [10]. Feature selection helped to identify the most relevant variables that contribute to accurate predictions. By incorporating CatBoost, a gradient boosting algorithm, this approach leveraged the power of ensemble learning and handles categorical features effectively. The application of this combined approach to above ground biomass estimation demonstrated its potential in accurately predicting biomass levels, which is valuable for environmental monitoring, forestry, and other related fields. This methodology showcased the effectiveness of combining feature selection and CatBoost for improved prediction accuracy in specific domains like above ground biomass estimation. [11] Introduced CatBoost, a gradient boosting algorithm specifically designed to handle categorical features effectively. Unlike traditional gradient boosting algorithms, CatBoost can directly handle categorical variables without the need for manual encoding or preprocessing. It employs a novel algorithm that incorporates statistical methods and gradient-based optimization to handle categorical features seamlessly. This capability makes CatBoost particularly useful in domains where categorical variables play a significant role, such as marketing, recommendation systems and credit scoring. By efficiently utilizing categorical features, CatBoost enhances predictive accuracy and performance of gradient boosting models, making it a valuable tool for various machine learning applications.

Prokhorenkova, Liudmila, et al. [12] explored a gradient boosting algorithm that addresses bias in boosting models when dealing with categorical features. Traditional boosting algorithms can introduce bias when handling categorical variables, leading to suboptimal predictions. CatBoost overcomes this challenge by employing an innovative algorithm that considers the statistical properties of categorical features. It effectively handles the categorical variables, reduces bias and improves the overall performance of gradient boosting models. This capability makes CatBoost a valuable tool for predictive modeling tasks that involve categorical features, offering unbiased predictions and enhancing the accuracy and reliability of the boosted models. Exploration of Amazon Web Services (AWS) features was done in [13].

When selecting a cloud service provider, it is essential to consider the distinguishing features of major platforms like AWS, Google Cloud Platform (GCP), and Microsoft Azure. AWS offers a wide range of services, extensive global infrastructure, and a mature ecosystem. GCP stands out with its advanced machine learning capabilities and seamless integration with other Google services. Microsoft Azure provides strong integration with Microsoft products, hybrid cloud solutions and robust support for enterprise workloads. Factors such as scalability, pricing models, storage options, security measures, and additional services must be evaluated to make an informed choice. Each platform has its strengths, and understanding these features helps businesses select a cloud service provider that aligns with their specific requirements. [14] Focused on the process of transitioning and deploying existing applications to the AWS cloud infrastructure. This migration involves various steps such as assessing application compatibility, selecting appropriate AWS services, designing architecture, and executing migration plan. By migrating applications to AWS, businesses can leverage the scalability, reliability and flexibility offered by cloud computing. This enables them to reduce infrastructure costs, improve performance, enhance security, and gain access to a wide range of AWS services. Successful migration to the AWS cloud empowers organizations to modernize their applications and unlock the benefits of cloud computing for their business operations. Varia and Jinesh [15] proposed a framework that aimed to extend differentiable machine learning (ML) pipelines beyond a single model. Traditionally, ML pipelines focus on optimizing individual models, but WindTunnel enables end-to-end optimization of complex pipelines. By incorporating differentiable components into the pipeline, such as feature extraction, preprocessing, and model selection, WindTunnel allows for seamless integration and joint optimization. This approach enhances the performance and efficiency of ML workflows by enabling automatic learning and adjustment of the entire pipeline. WindTunnel opens up possibilities for more comprehensive and effective ML pipelines that go beyond the limitations of optimizing individual models in isolation.

III. ARCHITECTURE

Proposed technical architecture of the software is depicted in figure 1. Application is deployed on AWS, particularly AWS Elastic Compute Cloud (EC2). Software is coded on Python. Web application is built using a framework called Streamlit. ML model is built using CatBoost algorithm and encapsulated using FastAPI. Web app connects with third party applications such as Mixpanel and GitHub; cloud services such as AWS Simple Storage Service (S3) Bucket to maintain functionality.

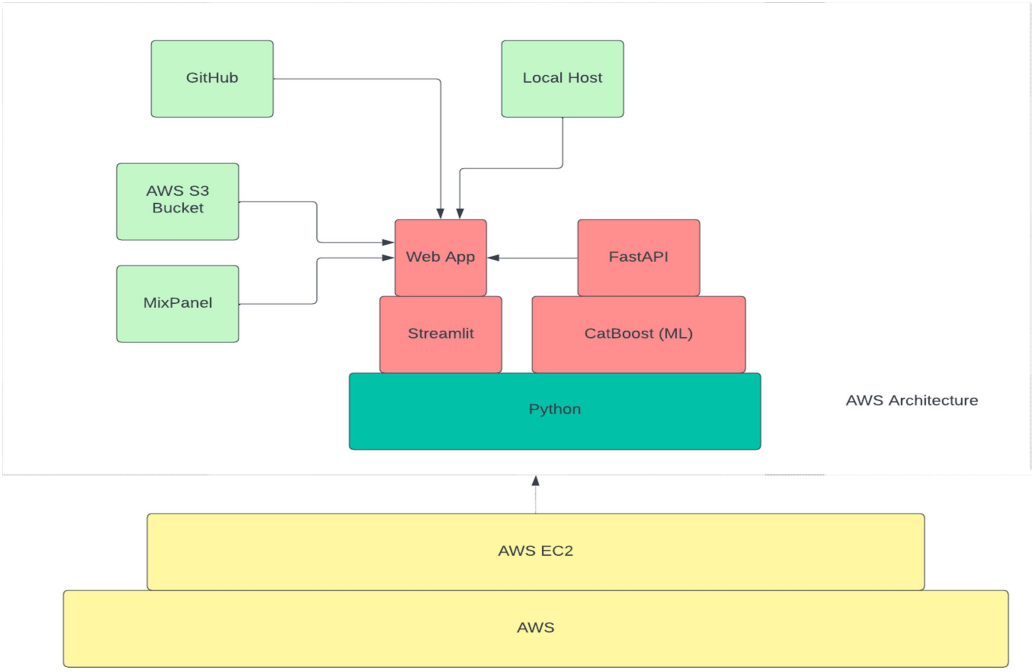
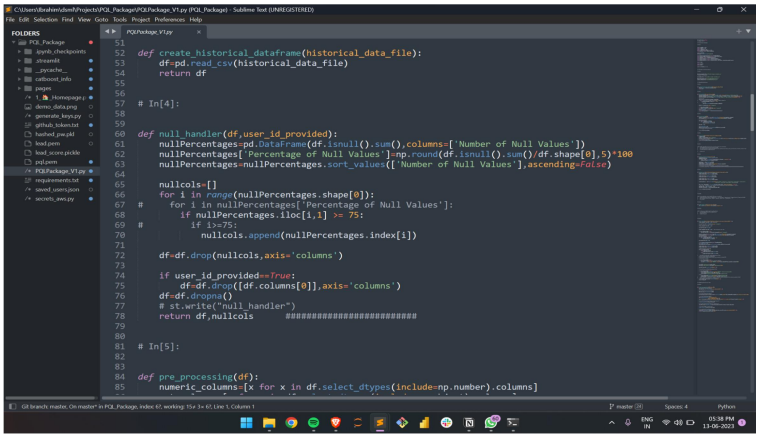


Fig. 1. Proposed Architecture

IV. METHODOLOGY

A. *Building ML Model*

A CatBoost algorithm is utilized to build an ML model. Utilization of CatBoost is motivated by its ability to handle categorical variables independently. A pipeline is constructed to encompass data cleaning, separation of numeric and categorical features, passing the data to CatBoost, and returning the results. An additional feature was implemented to generate the probability of a customer purchasing a product. This score will be employed as the lead score for a customer. The cut-off score will be determined by employing ROC and AUC curves which involved assessing the tradeoffs between False Positives and True Positives. Resulting cut-off score will be utilized subsequently for customer segmentation. The Model is encapsulated into a Python API using FastAPI. Figure 2 depicts the code design of ML model.



```
def create_historical_dataframe(historical_data_file):
    df=pd.read_csv(historical_data_file)
    return df


# In[4]:


def null_handler(df,user_id_provided):
    nullPercentages=pd.DataFrame(df.isnull().sum(),columns=['Number of Null Values'])
    nullPercentages['Percentage of Null Values']=np.round(df.isnull().sum()/df.shape[0],5)*100
    nullPercentages=nullPercentages.sort_values(['Number of Null Values'],ascending=False)

    nullcols=[]
    for i in range(nullPercentages.shape[0]):
#       for i in nullPercentages['Percentage of Null Values']:
#         if nullPercentages.iloc[1,1] >= 75:
#             if i>=75:
                nullcols.append(nullPercentages.index[i])

    df=df.drop(nullcols,axis='columns')

    if user_id_provided==True:
        df=df.drop([df.columns[0]],axis='columns')
    df=df.dropna()
    # st.write("null_handler")
    return df,nullcols      ##########################


# In[5]:


def pre_processing(df):
    numeric_columns=[x for x in df.select_dtypes(include=np.number).columns]
```

Fig. 2. ML Model

B. Building Web App

Decisions are made to utilize Streamlit for the development of web application. Streamlit is an open-source Python framework for constructing dynamic web apps, was selected. The web app is intended to comprise three sections: Home (Login), Stats, and Insights. Homepage will feature a login form where users will input their credentials provided by administrator. Login credentials will be hashed using Streamlit Login's built-in hash function. Figure 3 shows the homepage for users to enter their credentials.

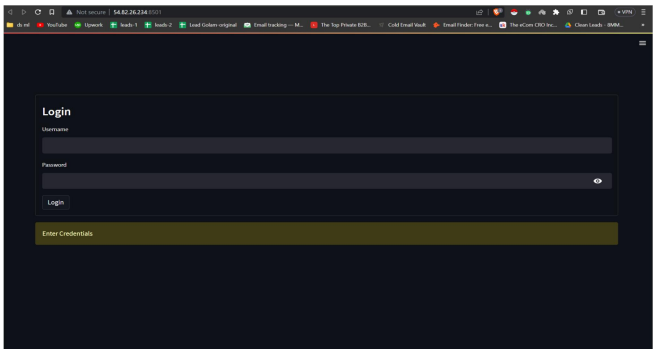


Fig. 3. Homepage

Subsequent step involves establishing a connection with a data source. The three primary sources of data for customers are local system, cloud system and CRM tool. To encompass these sources in our prototype, upload from PC, AWS S3 Bucket, and Mix- Panel are the methods employed. Additionally a section has been included to facilitate the utilization of demo data for educational purposes. Access to AWS Bucket and MixPanel is achieved using their respective APIs. Figure 4 illustrates the integration of data source.

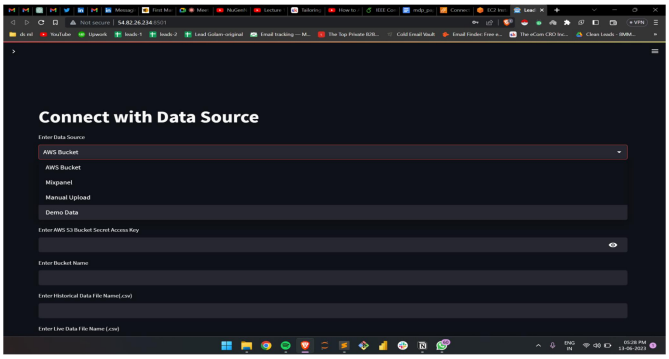


Fig. 4. Integrating Data Source

Stats page consist of segmenting the consumers based on their scores. On top left, it displays quantities of cold, warm and hot leads. On top right, a donut chart with the segmentation is displayed. The next section will contain top hot leads and warm leads based on their scores.

Insights page is specifically meant to analyze which type of users are not willing to buy the product or service and what are factors that affect their purchasing decisions. Clients can look at the deviations from an ideal customer. This will help the users decide which features actually differentiate a warm or cold lead from a hot lead. There is also a section where individual user id's can be entered and all feature variations can be analyzed. Figure 5,6 and 7 depicts the Stats - Customer Segmentation, Stats - Top Leads and Insights - Deviations from Ideal Values respectively.

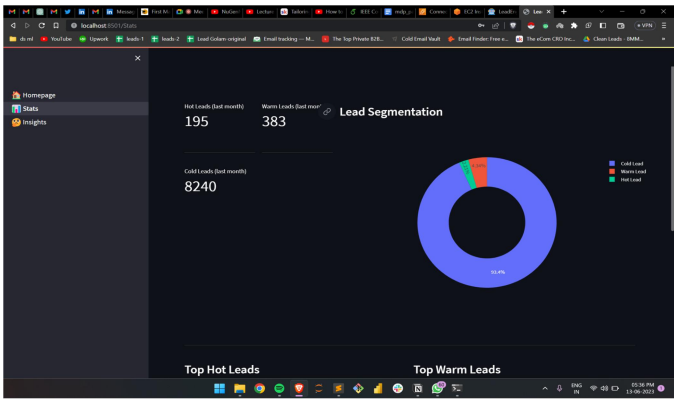


Fig. 5. Stats - Customer Segmentation

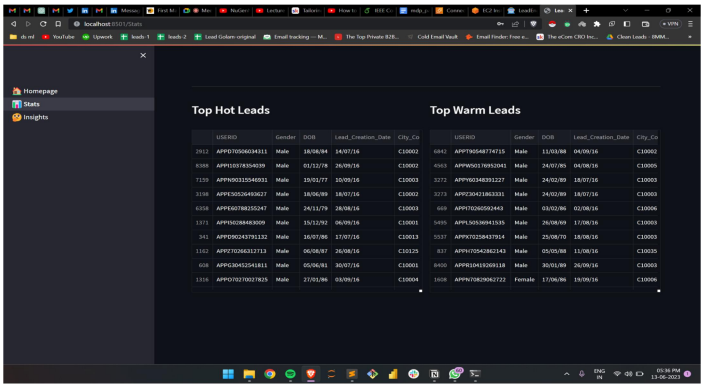


Fig. 6. Stats - Top Leads

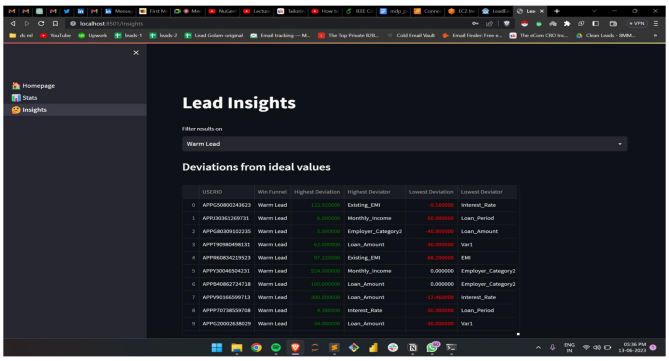


Fig. 7. Insights - Deviations from Ideal Values

Further improvements are to be made in providing call-to- actions, which can become a full stack CRM tool to be provided to clients on a subscription basis.

C. Deploying the web app on AWS

An account was created on Amazon Web Services to lever- age the benefits of free monthly credits and deploy the application. Initially, an AWS Elastic Compute Cloud (EC2) Instance is established. EC2 instance functions as a cloud-based development environment. AWS operates on a pay-as-you- use model, wherein users are billed for services they currently utilize.

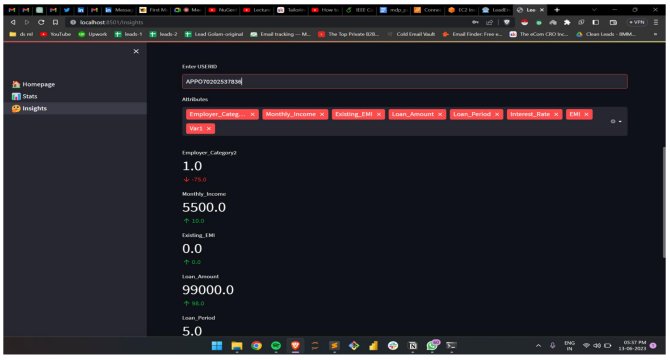


Fig. 8. Insights - Individual Lead Deviations

For experiment's instance, the t2. micro tier (free tier) and 8GB of storage were chosen. Following the creation of instance, a codebase repository is generated on GitHub which serves as a platform for sharing code among developers. Subsequently all Python dependencies were installed onto the instance. This could be accomplished either by running a virtual environment on local machine or utilizing built-in virtual environment called Instance Connect for managing installations. Each instance is assigned its own set of Public and Private IP Addresses, which act as temporary domain for the application. Initially GitHub repository is cloned within the instance, and dependencies were installed using pip (pip Install Packages). Pip is a command-line tool used for installing Python packages. Instead of manually downloading each dependency, requirements.txt file is employed to install all necessary packages at once. Finally, similar to running a Python application on the local host,

Streamlit web app is launched using the command “streamlit run appname”. Running the application redirects to a page displaying an IP address. Upon purchasing a domain name, this IP address could be mapped to it. Figure 8 and 9 shows Insights - Individual Lead Deviations and AWS EC2 Instance Connect respectively.

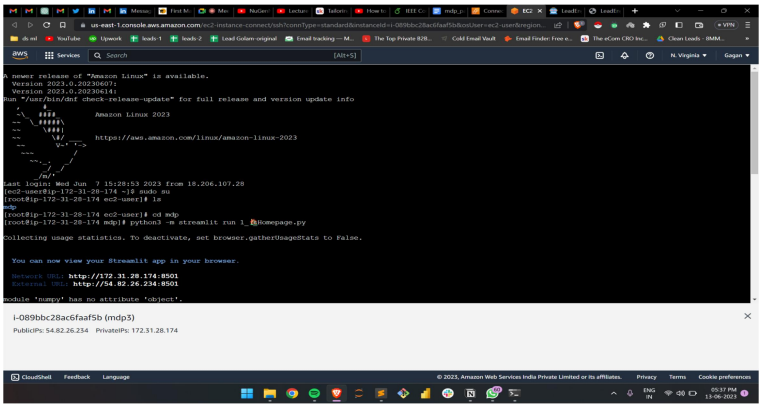


Fig. 9. AWS EC2 Instance Connect

V. FLOW OF PROPOSED SERVICE

User logs in to web app using credentials. First step after logging in is to integrate with a data source. Integration can be done in 3 ways, AWS S3 Bucket, Mixpanel or Localhost or user can opt for demo data. Next page profiles and segments users based on scores. Final page provides analysis and deviations through which data-backeddecisions to acquire customers can be made. Figure 10 illustrates the flow of proposed service.

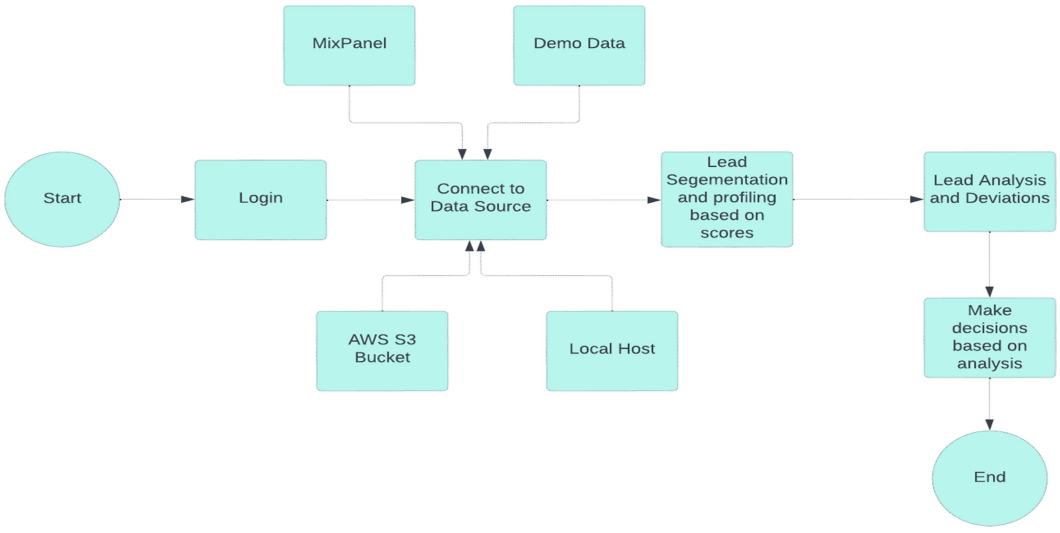


Fig. 10. Proposed Service for lead scoring

VI. CONCLUSION

Manual process of lead scoring is gradually diminishing, and a shift towards a data-first approach is attracting companies. While this micro-service offers a simple functionality, ultimate vision is to develop an end-to-end CRM SaaS solution that aids sales and marketing divisions in reducing the time spent on customer acquisition. ML model can be optimized to incorporate online learning techniques, such as batch or streaming processing to enable retraining. Additionally CatBoost algorithm can be optimized for better alignment with the sales objectives. API can also incorporate PyCaret, a tool that assists in selecting the most suitable ML algorithm for a given dataset and facilitates comparison between algorithms. PyCaret is widely used for Automated Machine Learning (AutoML) allowing for quick analysis of various algorithms' performance and aiding in decision-making in choosing best algorithms for the data.

VII. ACKNOWLEDGEMENTS

We thank the institute, B.M.S. College of Engineering for the wonderful opportunity of learning. We extend our gratitude to the department of Information Science and Engineering for facilitating the learning process.

REFERENCES

[1] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system" Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[2] Dordevic, Aleksandar. "Optimization of digital marketing processes through modeling of lead-scoring" Sinteza 2019-International Scientific Conference on Information Technology and Data Related Research.
Singidunum University, 2019

[3] Benhaddou, Youssef, and Philippe Leray. "Customer relationship man- agement and small data—Application of Bayesian network elicitation techniques for building a lead scoring model" 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). IEEE, 2017.

[4] Slakey, Austin, Daniel Salas, and Yoni Schamroth. "Encoding categor- ical variables with conjugate bayesian models for wework lead scoring engine" arXiv preprint arXiv:1904.13001 (2019).

[5] Nygård, Robert. "AI-Assisted Lead Scoring" (2019).

[6] Etminan, Ali. "Prediction of Lead Conversion with Imbalanced Data: A method based on Predictive Lead Scoring." (2021).

[7] Swelsen, CWJM Caro, et al. "Proposing a Generic Online Lead Scoring Model for a B2C Market" (2019).

[8] Naveen Kumar, G., and K. Hariharanath. "Designing a Lead Score Model for Digital Marketing Firms in Education Vertical in India" Indian Journal of Science and Technology 14.16 (2021): 1302-1309.

[9] Jadli, Aissam, et al. "Toward a Smart Lead Scoring System using Machine Learning" Indian Journal of Computer Science and Engineering 13.2 (2022): 433-443.

[10] Luo, Mi, et al. "Combination of feature selection and catboost for pre- diction: The first application to the estimation of aboveground biomass" Forests 12.2 (2021): 216.

[11] Dorogush, Anna Veronika, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support" arXiv preprint arXiv:1810.11363 (2018).

[12] Prokhorenkova, Liudmila, et al. "CatBoost: unbiased boosting with categorical features" Advances in neural information processing systems 31 (2018).

[13] Kamal, Muhammad Ayoub, et al. “Highlight the features of AWS, GCP and Microsoft Azure that have an impact when choosing a cloud service provider” Int. J. Recent Technol. Eng 8.5 (2020): 4124-4232.
[14] Varia, Jinesh. ”Migrating your existing applications to the aws cloud. “ A Phase-driven Approach to Cloud Migration” (2010): 1-23.
[15] Yu, Gyeong-In, et al. “WindTunnel: towards differentiable ML pipelines beyond a single model” Proceedings of the VLDB Endowment 15.1 (2021): 11-20.