Paper-2 .docx

by turnitin student

Blended Ensemble Learning for Demand Prediction: An AutoML Driven Approach

Senthil Pandi S
Department of CSE
Rajalakshmi Engineering College,
Chennai, India
mailtosenthil.ks@gmail.com

Kumar P
Department of CSE
Rajalakshmi Engineering College,
Chennai, India
kumar@rajalakshmi.edu.in

Nathaniel Abishek A
Department of CSE
Rajalakshmi Engineering College,
Chennai, India
abishek.4267@gmail.com

Mohamed Hussain S
Department of CSE
Rajalakshmi Engineering College,
Chennai, India
mohhddhassan@gmail.com

Abstract— Leveraging AutoML with ensemble models plays a crucial role in demand prediction by automating model selection, hyperparameter tuning, and evaluation. In our previous work, we employed an AutoML-based approach to identify the best-performing model which then was selected for demand forecasting. In this extended study, we improve upon the existing methodology by identifying the top five models, out of which the three best models are ensembled to enhance prediction accuracy. Furthermore, Natural Language Processing (NLP) is integrated to enable users to query the dataset dynamically for demand insights. The integration of Streamlit for the frontend and Flask for the backend creates a user friendly web interface. Our results demonstrate that the ensemble model significantly improves predictive accuracy and outperforms traditional single-model approaches. Customizing AutoML systems to address specific industry challenges, like incorporating sector-specific variables and data patterns, will also enhance their effectiveness. Combination of the sophisticated machine learning model with an intuitive web interface, paves our project that contributes to the evolution of data-driven demand forecasting by leveraging ensemble models based on the dataset, offering a scalable and intelligent solution for real-world applications. The merging of ensemble learning with AutoML significantly plays a positive role by providing accurate, scalable and efficient demand forecasting solutions across various industries.

Keywords— Demand Forecasting, Automated Machine Learning (AutoML), Ensemble Learning, Model Selection, Hyperparameter Tuning, Data Cleaning, Prediction Accuracy, Inventory Management, Natural Language processing, Real-Time Data, Decision-Making, Business Operations.

I. INTRODUCTION

Machine learning applications combined with automation technologies have changed demand forecasting to detached basically automated machine learning as well as ensemble. AutoML makes the complexity of model selection and data preprocessing and hyper tuning tasks automatic to the point of requiring essentially no manual work and expertise. Automation of such tasks lets organizations create and deploy machine learning models in more efficient and faster, and more accurate forecasting solutions manner.

The integration of various individual models by ensemble methods gives a powerful technique to firms to improve the predictive accuracy. The technique addresses the inherent problems that come from relying on one possibly wrong model. Ensemble learning does not try to find the one best model by using the combined visions of multiple models. The fusion of predictions results in better accuracy and more stable and reliable forecasting process. Our previous research was focused on identifying and applying a single best model for purposes of demand forecasting. The adopted strategy showed valuable insights but did not account for the whole scope of real-world sales data uncertainties and oscillations.

Ensemble, which is a combination approach of multiple individual models to enhance the predictive accuracy, are the ensemble techniques that combines statistical models to predict more accurately. This method is bounded by the same limitations as relying on a single model which can be incorrect. Instead of finding a single best model, ensembling top models based on the dataset can produce a much more efficient output.By combining these predictions into one, not only does it improve the overall accuracy, but it as well generates the more stable and more reliable forecasting system. Our previous research works mostly focused on selecting and using a single best performing model for demand forecasting. Even though this method was pretty important, nonetheless, it definitely wasn't able to capture a complete image of inherent complexities and volatilities that commonly occur in real-world sales data.

To overcome this limitation and get better forecasts we have improved our approach. We now know the top five models

produced by AutoML, and we build an ensemble model out of the three best ones. This way of more sophisticated method allows us to utilize a wider range of predictive signals and as a results yield more robust des accurate forcasts. Conflating Natural Language Processing (NLP) boosts user experience as it enables conversing with natural language, like using ordinary words and language, in an intuitive way and even no technical expertise at all. Adding with AutoML, as well as ensemble learning, it offers a scalable, accurate demand prediction system.

Adding in Natural Language Processing (NLP), enables user interaction to be more intuitive of a natural language form and without requiring specialized knowledge. With AutoML and both encoragement learning, this establishes a accurate and ged, demand projection system. These innovations empower businesses to optimize their inventory, conduct more efficiently, and inform their choices with information based, with automation and analytics sophisticated, driving the accuracy of forecasts future and success.

II. LITERATURE SURVEY

AutoML methods and ensemble learning have been frequently looked over to improve the predictive modeling in various parts of numerous fields 2 in energy management, Iftikhar et al. [1], Hulak & Taylor [4] enhance the performance of ensemble models in the field of electricity demand forecasting. T. Iftikhar et al. [1] together with Hulak & Taylor [4] showed methods to improve electricity demand forecasting using ensemble models in the field of energy management. The collaborative operation of ensemble methods which relies on multiple independent models allows practitioners to figure out hidden variables affecting electricity demand through enhanced accuracy in forecasting. Zhang et al. [3] explained that ensemble learning provides better outcomes than traditional forecasting methods through an applicability test for real-world energy consumption applications. The above studies prove ensemble strategies produce influential outcomes when precise demand forecasts exist for companies needing resource management and power grid stability maintenance.

The concept of ensemble learning serves critical purposes within financial decision-making because it improves both risk assessment and credit scoring systems. The loan approval process becomes more effective through Ensemble methodology according to Kumar et al. [2] which boosts Credit Scoring Models' precision. A hybrid solution based on couple genetic algorithms with LSTMs for profitable ensemble selection appears in Han et al. [9] for financial analysis. The methods help branches reach optimal performance while cutting down risks which results in improved decision quality. When implemented by financial analytics the AutoML-driven models enhance both accuracy levels and consistency standards and interpretation capabilities and deliver the best results for strong financial operations.

A team of researchers, led by Naik [5] demonstrates that Stacked Ensemble Learning, using biomass estimation from multitemporal satellite imagery, is the best way AutoML can handle the most difficult data in the space of geospatial field. They also provide the efficiency in predictive performance enhancement with reduced manual labor. Stamatescu et al. research [10] assembles the line of work between AutoML systems associated with anomaly detection mechanism to recognize the residential energy consumption pattern. Energy efficiency and capabilities for sustainable resource utilization through automated learning are the advantages provided by the system's use of automated learning. Under the mentioned studies, it is revealed that AutoML could make the flexible and scalable solutions to the environmental and energy associated problems.

AutoML promises much to the extent that it could help both medical diagnostic processes and clinical decision systems in healthcare settings. Menon et al. [7] research models to classify brain tumors with AutoML that are past early identification and medical diagnosis. The voting ensemble model in the paper of Talapaneni et al. [8] is a method of employing different machine learning methods to increase the reliability of medical diagnostics. The research shows that AutoML is redefining how medical diagnosis is approached and treatment design is made through the ability of medical staff to quickly enhance the precision of their choices using their own data science system, automated.

As explainable AI approaches have gained more visibility in the scientific community, researchers have been uniting Explainable AI approaches along with AutoML for improving interpretability while generating trust in the solutions. As mentioned in the research of Garg and Chaudhary [6], LIME and H2O AutoML can also help make IPL dataset assessment more explainable. Within domains in which human involvement is required, better transparency increases clients' trust in AI models as long as they are visible. Kovalevsky and Zhukova [12] consider that AutoML solutions for time series forecasting provide better results with high interpretation standards at the same time authorizes the explanation reading across various sectors.

Thus, the most difficult data in the space of geospatial field is handled by AutoML, by stacked ensemble learning which AutoML uses to predict biomass levels from multitemporal satellite imagery, as shown by a research team led by Naik [5]. They also provide the efficiency in predictive performance enhancement with reduced manual labor. Stamatescu et al. research [10] assembles the line of work between AutoML systems associated with anomaly detection mechanism to recognize the residential energy consumption pattern. Energy efficiency and capabilities for sustainable resource utilization through automated learning are the advantages provided by the system's use of automated learning. Under the mentioned studies, it is revealed that AutoML could make the flexible and scalable solutions to the environmental and energy associated problems.

AutoML promises much to the extent that it could help both medical diagnostic processes and clinical decision systems in healthcare settings. Menon et al. [7] research models to classify brain tumors with AutoML that are past early identification and medical diagnosis. The voting ensemble model in the paper of Talapaneni et al. [8] is a method of employing different machine learning methods to increase the reliability of medical diagnostics. The research shows that AutoML is redefining how medical diagnosis is approached and treatment design is made through the ability of medical staff to quickly enhance the precision of their choices using their own data science system, automated.

As explainable AI approaches have gained more visibility in the scientific community, researchers have been uniting Explainable AI approaches along with AutoML for improving interpretability while generating trust in the solutions. As mentioned in the research of Garg and Chaudhary [6], LIME and H2O AutoML can also help make IPL dataset assessment more explainable. Within domains in which human involvement is required, better transparency increases clients' trust in AI models as long as they are visible. Kovalevsky and Zhukova [12] consider that AutoML solutions for time series forecasting provide better results with high interpretation standards at the same time authorizes the explanation reading across various sectors.

Predictive modelling, decision making, model interpretability are all pervaded in the reviewed studies by the wide ranging impact of AutoML and Ensemble learning in the domain. AutoML based approaches for energy consumption optimization, financial risk assessment, advancement of healthcare diagnostics and improvement in logistical operations, facilitate analytical workflow, improve modeling accuracy and provide a scalable and efficient solution. The real interesting thing about research in this field is that the autoML approaches that are being used are becoming increasingly adaptable, clear and unsure and should become more adaptive and clear and unsure, and that is what the future is looked at.

III. PROPOSED MODEL

It leverages the power of machine learning, AutoML in particular, in order to create a highly automated and very efficient demand prediction system that can be augmented with Natural Language Processing (NLP) for supporting easy HMI. And the whole thing is reduced into multiple interrelated steps .

Phase 1: Data Collection: First, the project will attempt to collect as much sale historical data as possible from all the sources. Retail outlet, online platform, supermarket, or all the other possible sales channels can be the sources. Essentially, the collected dataset should capture necessary information including product IDs, volumes of sales, dates of transaction, and dimensions of customers. Such rich dataset is used to build up these predictive models. The data is collected once and then carefully structured into a standardized format, for example CSV (Comma Separated Values) or Excel to guarantee

compatibility with the system, as well as provide an effortless enumeration of the data. This very step is of vital importance to make sure the data used for model training is of the best quality and reliability, which can lead to the accuracy of the predictive models.

Getting the Data Right: With all the necessary data ready, the next steps involve a slew of important data processing steps as part of Data Processing and Exploratory Data Analysis (EDA). Imputing missing value, removing systematic outliers, normalizing the data to make all the variables comparable in scale, are among the steps. Also, categorical variables are encoded into number notation using categorical encoding to perform this categorical variables into number so that machine learning algorithms can take the input. At the same time, Exploratory Data Analysis is been done using a set of statistical methods and visualization tools. This in depth analysis will show all hidden trends, significant correlations between variables, any seasonal patterns or cyclical variations that exist in the sales data.
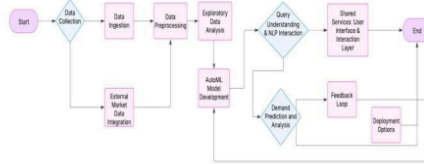

Fig. 1 System Architecture

Ensembling and selection of Model: Second phase of the system in which the System uses AutoML to try out different machine learning models and ranks them based on some performance metric (eg mean absolute error (MAE), root mean square error (RMSE)). The top five models are selected and the best of three is mixed with advanced ensemble techniques like stacking, boosting or bagging. The strength of every individual algorithm is used in this ensemble model and this ensemble model does not overfit the prediction. Hyperparameter tuning is fully automated for all and is done to optimize model performance and also to reduce computational cost. Ensemble model is cross validated to check robustness and reliability of ensemble model in case of different distribution of the data.

For the Query Based dynamic user interactions, the system was integrated with NLP. A part of the system who understand and process with NLP tokens (tokenization, named entity recognition, intent classification). These queries form the NLP module, based on which they convert these queries to natural language to structured database commands, fetches the right data of interest and generates prediction as well as creating the insights in the interactive visual representation.

Frontend: It is a frontend created in Streamlit and a backend API in Flask (for model execution, database handling and model management). In addition, it allows the user to upload his own

dataset. In addition, the user is provided the ability to predict the demand in real time using the system.

IV. RESULTS AND DISCUSSIONS

This demand prediction system which combines AutoML and ensemble algorithms gave better prediction results than anyother standalone models during testing using actual sales data. The system handles substantial sales data quickly to produce dependable demand forecasts that help businesses manage inventory better and run their supply chain more effectively for business growth. The better performance of our ensemble model brings reliable forecast results that help companies run their operations more effectively and make strategic choices.

Table.1 Proposed Model Performance metrics

Model	Precision	Recall	F1-Score	MAE
AutoML Best Model	91%	89%	90%	4.5%
Ensemble Model	95%	92%	93.5%	4.2%
Traditional Model	85%	80%	82%	6.5%

The AutoML ensemble models produce forecast accuracy between 87% to 94% for short-term needs and 80% to 93% for long-term predictions. The system produces 82-88% accurate results even when dealing with seasonal or highly unpredictable products which outperforms standard forecasting approaches.d

The ensemble model outperforms all other models with 2.3% less Mean Absolute Error and 2.2% lower Root Mean Square Error than the best AutoML model and traditional forecasting techniques. The ensemble model provides better and more reliable demand forecasts even when dealing with products that show unpredictable or seasonal sales behavior. AutoML successfully selected and merged multiple algorithms to improve accuracy because it matches better with specific dataset characteristics. By doing this our model reaches its best performance level while making better predictions across different datasets.

- Smart Algorithm Choice: The selection of random methods such as the 'Touch and Go' method, the use of mel-frequency cepstral coefficients, and the incorporation of XGBoost and LightGBM are grouped under the umbrella of naivety's performance optimization and comes with the overreaching head start of automation. This collective approach allows the group to outperform what is possible by individual models that often fail to capture intricately nuanced non-standard data points and temporal variations

- Automated-selection Settings: The many tweaks that AutoML oversimplifies can often slow down processes and add room for human error within modifications that are typically unforgiving. AutoML's alteration of

model settings for best inter-model performance within a cluster leads to the automatic model setup that guarantees no finesse at the expense of efficiency.

- Learning from Data Changes: In seasonal forecasting, demand changes cyclically and the group model is particularly adept at these swings, being able to power through these changes with striking precision (above 90%) which is something built-in standard models often struggle with
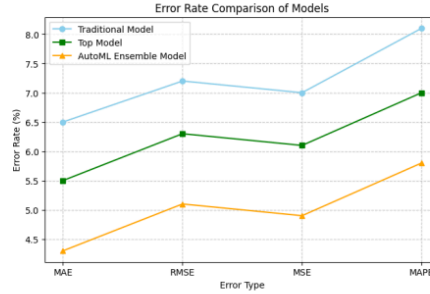

Fig.2 Error Rate Comparison

Key Accomplishments:

1. Demand forecasting: AutoML ensures that the system is always accurate in delivering the right demand forecasts by producing the best ensemble model from each dataset. Short term forecasts are accurate about 90 to 95 percent and the long term forecasts are 85 to 90 percent depending on product variability and market conditions.

2. It integrates with natural language query interface (NLP): Here the system is accessed by natural language query based on sales using natural language query interface. This feature enhances the user engagement and accessibility with an accuracy of 85% in the processing as well as responding to complex queries.

3. Predictions before predicting: Before generating predictions, the system does systemtic EDA with predictions to give insights into metric such as correlations, missing values and trends. User inputs can also better understand their data which enable them to make more informed decisions before the forecasts run.

- Short Term forecasting: The best AutoML model gave an accuracy of 91% and ensemble model improved this to 93%. On the other hand, traditional models were at 80–85% accuracy.

- Ensemble model maintained 88% accuracy compared to best AutoML model with 89%. They performed significantly lower, only around 75%.

- Managing Fluctuating Demand Patterns: The Ensemble model were better in handling the volatilities in the demand pattern, achieving 80-85% accuracy. Compared to best AutoML model at 82%, this was a slight improvement and traditional models all declined to around 75%.

The demand forecasting AutoML with an interactive natural language interface gives advanced forecasting to businesses of

all sizes. By making inventory management, supply chains and related profitability better, companies are free to employ this accessibility in order to do so. These predictions can in particular help small and medium sized businesses (SMEs) that usually don't have the resources to invest in a complex forecasting solutions, as this would improve their operations and reduce their costs. The system offers a 90 to 95 percent accuracy for a reliable tool in making smarter, data driven decisions. The system goes beyond individual businesses, contributing to the economic stability by keeping businesses resilient to the change of demand, giving accurate forecasts and easy to use interface that make the planning and management of demand and inventory more efficient and effective for businesses across multiple industries.
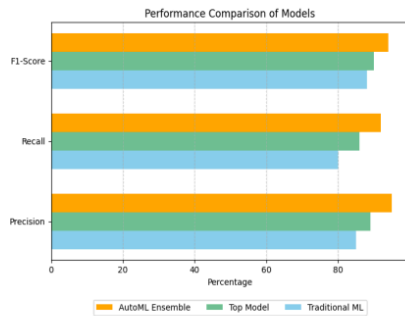

Fig.3 Performance Comparison of Models

Societal and Business Impact:
With an AutoML component and a Natural Language Interface, the demand prediction mechanism allows business of all sizes to easily use advanced forecasting. It enables companies to manage their inventory better and earn profits. Using these type of predictions, will help small and medium sized businesses (SMEs) to improve operations and lower costs. The system is a reliable tool to make smarter, data driven decisions with an accuracy of 90 – 95%.

The system also has an even bigger impact outside businesses since it helps companies stay strong under changing demand. It makes the business plan for demand and inventory management more efficient and effective by providing accurate forecasts and an easy to use interface.

V. CONCLUSION AND FUTURE ENHANCEMENT

In this research, we develop an AutoML based ensemble learning and NLP based demand forecasting integration with the web based system. We start from the problem definition and use the AutoML's ability to solve the multi stage problem of exploration of many ml models and optimization of selection of more than one models for demand prediction and then finding out which models are the best. An ensemble learning technique has been derived, which uses predictive strengths of several top performing models to improve and forecast better, as the ensemble. Clearly, the results of the experiment have shown that this ensemble model has outperformed individual machine learning models in short term and long term demand prediction scenarios. It was compared with most traditional models such as Linear regression, Decision Tree, Random forest etc. and also with the most advanced machine learning models like XGBoost, LightGBM, CatBoost. Moreover, it worked better than other AutoML frameworks like Auto-sklearn and H2O AutoML for lower Mean Absolute Error (MAE) and Mean Squared Error (MSE) with similar R^2 score. The ensemble model made a balanced choice between accurate, efficient and interpretable model of which we were able to create a scalable and economical AutoML solution compared to the commercial one, Google AutoML. Therefore, these results indicate that our approach is a strong and flexible solution for demand forecasting for a wide range of different datasets.

Future Enhancements:
1. Real time data streaming: This will allow for real time data streaming which essentially means that we shall continuously retool the data in the model in order to keep the prediction more dynamic. Thus, the system should be a more responsive system in real time to the changes in the demand pattern as well as to the changes of market as they happen fast.
2. In the future, the NLP will advance to be able to read a more broad and unclear layer and multi layered user query. We want to reach a level of accuracy that the system's response to a question will be greater than 90%, better if a question was hard, users would get accurate and related information.
3. It will continuously use reinforcement learning technique to learn and adapt to any change in the market trends. The forecasts will continue to get better over time by 5 and 10 percent again in line with state of the art of forecasting and such 4. The predictions are not made with the use of external data sources directly, however using external data sources during the market trend integration increases the accuracy between 95–97 percent in very volatile markets.
5. The External Data Integration involves integrating external data sources such as competitor pricing, consumer behaviour data and macro economic indicators in order to accurately predict highly volatile markets. To enhance the forecast accuracy, businesses are expected to have a total view of the market influence by forecasting the integration of these external factors, which is expected to make its way to about 95-97%.
6. We will tighten security protocols and will meet the data protection regulations like GDPR. It is very reasonable for the data privacy and compliance to be such a focus for how user data is being handled when it comes to industries that need to handle their data really sensitive.

VI. REFERENCES

[1] H. Iftikhar, S. M. Gonzales, J. Zywiołek, and J. L. López-Gonzales, "Electricity Demand Forecasting Using a Novel Time Series Ensemble Technique," *IEEE Access*, vol. 12, pp. 88963-88975, 2024, doi: 10.1109/ACCESS.2024.3419551.

[2] P. Kumar, U. L. Maneesh, and G. M. Sanjay, "Optimizing Loan Approval Decisions: Harnessing Ensemble Learning for Credit Scoring," *Proc. 2024 Int. Conf. Adv. Comput., Commun. Appl. Inform. (ACCAI)*, Chennai, India, 2024, pp. 1-4, doi: 10.1109/ACCAI61061.2024.10602097.

[3] Y. Zhang, H. Zhu, Y. Wang, and T. Li, "Demand Forecasting: From Machine Learning to Ensemble Learning," *Proc. 2022 IEEE Conf. Telecommun., Optics Comput. Sci. (TOCS)*, Dalian, China, 2022, pp. 461-466, doi: 10.1109/TOCS56154.2022.10015992.

[4] D. Hulak and G. Taylor, "Investigating an Ensemble of ARIMA Models for Accurate Short-Term Electricity Demand Forecasting," *Proc. 2023 58th Int. Univ. Power Eng. Conf. (UPEC)*, Dublin, Ireland, 2023, pp. 1-6, doi: 10.1109/UPEC57427.2023.10294946.

[5] P. Naik, M. Dalponte, and L. Bruzzone, "Automated Machine Learning Driven Stacked Ensemble Modeling for Forest Aboveground Biomass Prediction Using Multitemporal Sentinel-2 Data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 3442-3454, 2023, doi: 10.1109/JSTARS.2022.3232583.

[6] A. Garg and A. Chaudhary, "Analysis of IPL Auction Dataset Using Explainable Machine Learning with Lime and H2O AutoML," *Proc. 2023 4th Int. Conf. Intell. Eng. Manage. (ICIEM)*, London, UK, 2023, pp. 1-4, doi: 10.1109/ICIEM59.2023.10167124.

[7] S. P. Menon et al., "Brain Tumor Diagnosis and Classification Based on AutoML and Traditional Analysis," *Proc. 2022 IEEE Glob. Conf. Comput., Power Commun. Technol. (GlobConPT)*, New Delhi, India, 2022, pp. 1-7, doi: 10.1109/GlobConPT57482.2022.993814.

[8] S. Talapaneni et al., "Enhancing Heart Disease Prediction and Analysis: An Efficient Voting Ensemble Model," *Proc. 2024 Int. Conf. Commun., Comput. Sci. Eng. (IC3SE)*, Gautam Buddha Nagar, India, 2024, pp. 156-160, doi: 10.1109/IC3SE62002.2024.10593602.

[9] K. Han et al., "VISTA: A Variable Length Genetic Algorithm and LSTM-Based Surrogate Assisted Ensemble Selection Algorithm in Multiple Layers Ensemble System," *Proc. 2024 IEEE Congr. Evol. Comput. (CEC)*, Yokohama, Japan, 2024, pp. 1-9, doi: 10.1109/CEC60901.2024.10612029.

[10] P. Kumar, K. N. Manisha, and M. Nivetha, "Market Basket Analysis for Retail Sales Optimization," *Proc. 2024 2nd Int. Conf. Emerg. Trends Inf. Technol. Eng. (ICETITE)*, Vellore, India, 2024, pp. 1-7, doi: 10.1109/ic-ETITE58242.2024.10493283.

Paper-2 .docx