# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Aibarna Singh Basnet |
| **Project Name** | AT2: Machine Learning as a Service (Retail) |
| **Date** | 6/10/2023 |
| **Deliverables** | * [Prediction Model](https://github.com/aibarna/basnet_aibarna-24585717-ML_AT2/blob/main/notebooks/predictive/basnet_aibarna_24585717_prediction_lgbm.ipynb) * [Forecast Model](https://github.com/aibarna/basnet_aibarna-24585717-ML_AT2/blob/main/notebooks/forecasting/basnet_aibarna_24585717_forecasting_arima.ipynb) * [Github Repo](https://github.com/aibarna/basnet_aibarna-24585717-ML_AT2/tree/main) * Prediction model on Heroku * Forecasting model on Heroku |

|  |  |
| --- | --- |
| 1. **EXPERIMENT BACKGROUND** | |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | |
| **1.a. Business Objective** | Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?  The goal of this project for the business is to create a predictive and forecasting model for sales, enabling the optimization of inventory management, staff allocation, and overall store performance across California, Texas, and Wisconsin. Accurate sales forecasts will guide decisions in inventory management, ensuring products are neither overstocked nor understocked. It will also inform staff scheduling, aligning labor hours with expected customer demand. By tailoring strategies to well-performing and underperforming stores, the project aims to boost overall sales and enhance customer satisfaction. The results will be instrumental in refining marketing strategies, assessing their ROI, and improving the customer experience.  The impact of accurate results is substantial, leading to cost savings, increased revenue, enhanced customer satisfaction, and efficient resource allocation. It allows for streamlined operations and well-planned marketing efforts, ultimately leading to a more profitable and customer-centric business. In contrast, incorrect results can result in overstocking, stockouts, inefficient staff allocation, and missed revenue opportunities, all of which can have detrimental effects on costs, revenue, and customer satisfaction. Therefore, the accuracy of sales predictions is critical to the business's overall success and profitability. |
| **1.b. Hypothesis** | Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it,  Hypothesis:  The hypothesis we aim to test in this project is whether the seasonality and sales patterns observed in the retail stores in California (CA), Texas (TX), and Wisconsin (WI) significantly differ. Specifically, we want to understand if sales in these states exhibit distinct seasonal trends and if these trends impact store performance differently.  Reasons for Considering this Hypothesis:  1. Geographical Diversity: CA, TX, and WI represent diverse regions with varying climates and consumer preferences. Testing this hypothesis will help us gain insights into how geographical factors affect sales trends.  2. Operational Decision-Making: Understanding seasonal variations is crucial for inventory management, marketing, and staffing decisions. Accurate insights can lead to cost savings and increased revenue.  3. Regional Customization: Different sales patterns in each state may call for tailored strategies. By validating this hypothesis, we can refine our approaches to better meet local customer demands.  4. Efficient Resource Allocation: Recognizing seasonal trends allows for the efficient allocation of resources, reducing unnecessary costs and enhancing overall store performance.  By testing this hypothesis, we seek to make data-driven decisions and harness the power of seasonality insights to improve business outcomes. |
| **1.c. Experiment Objective** | Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.  Expected Outcome:  The expected outcome of the experiment is to uncover and validate the presence of distinct seasonal sales patterns in retail stores across California (CA), Texas (TX), and Wisconsin (WI). We anticipate observing variations in sales trends between these states due to differences in climate, regional customs, and customer preferences.  Estimate of the Goal:  While the specific quantitative goals may vary, the overarching goal is to gain a deep understanding of seasonal sales trends in each state. This understanding will enable us to:  1. Identify key seasonal periods when sales are at their peak or lowest.  2. Tailor inventory management to match demand during these periods, optimizing stock levels.  3. Adjust marketing strategies to capitalize on regional seasonal preferences.  4. Efficiently allocate staff based on anticipated customer footfall.  Possible Scenarios:  1. Validation of Distinct Seasonal Trends: The experiment could confirm the hypothesis, showing that CA, TX, and WI have unique sales patterns influenced by regional factors. This would be the most favorable outcome, allowing for precise season-based strategies.  2. Partial Confirmation: It's possible that while some distinct seasonal patterns are observed, others are similar across the states. In this case, we can still make informed decisions but with a degree of overlap in strategies.  3. No Significant Difference: The experiment might reveal that the sales patterns are largely uniform across the states. While this would simplify operations, it might limit the potential for region-specific optimizations.  4. Unexpected Findings: Occasionally, experiments yield surprising insights. There's a chance of uncovering entirely unexpected sales trends that challenge conventional wisdom.  The aim is to obtain insights that empower the business to make data-driven decisions in line with the observed seasonal trends, ultimately enhancing operational efficiency and customer satisfaction. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS** | |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | |
| **2.a. Data Preparation** | Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments  Data Preparation Steps Taken:  1. Data Collection: Comprehensive data was gathered from internal sources, ensuring accuracy and completeness. Multiple data sets encompassing store information, sales data, inventory levels, customer behavior, geographical distribution, marketing data, and seasonal sales patterns were collected.  2. Data Cleansing: Data underwent thorough cleaning to identify and rectify missing values, errors, and inconsistencies. This step was crucial to ensure that the data used for analysis is accurate and reliable. Missing information was filled in, and erroneous entries were removed.  3. Data Integration: Data from diverse sources and formats were integrated into a unified dataset for analysis. This step allowed us to work with a single, comprehensive dataset, simplifying the analysis process.  4. Exploratory Data Analysis (EDA): Basic statistics and visualizations were employed to gain an in-depth understanding of the data. Summary statistics, histograms, scatter plots, and other visualizations unveiled patterns and trends in the data.  Rationale:  These data preparation steps were essential to ensure the quality and integrity of the data used for analysis. By cleaning the data and integrating it into a unified dataset, we could perform meaningful analyses and derive valuable insights. EDA helped us understand the data's characteristics and identify key trends, which will inform subsequent analyses and decision-making.  Steps Not Executed and the Reasoning:  1. Advanced Feature Engineering: While basic EDA was conducted, more advanced feature engineering was not performed in this phase. The rationale was to first gain a comprehensive understanding of the data's existing features and characteristics. Advanced feature engineering may be considered in future experiments to create more sophisticated predictive models.  Important for Future Experiments:  Data integration and EDA are crucial steps that will set the foundation for future experiments. They enable us to work with high-quality data and extract valuable insights. In subsequent experiments, advanced feature engineering, if necessary, could enhance the predictive models' performance and accuracy. |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments  The project involved several steps for generating features, each driven by specific rationales:  1. Store Information Features: Features pertaining to store details, including location, size, and store type (e.g., flagship, outlet), were generated. These features were crucial to contextualize the sales data and enable the segmentation and analysis of store performance based on these attributes.  2. Sales Data Features: A wide range of sales-related features were created, including daily, weekly, and monthly sales figures for individual stores, product categories, and overall total revenue. These features were vital for understanding sales performance, identifying popular products or product categories, and recognizing sales trends over time.  3. Inventory Management Features: Features that captured inventory-related information, such as the rate at which products are sold and the availability of popular items, were generated. These features were instrumental for efficient inventory management, ensuring that products were stocked optimally.  4. Customer Behavior Features: Features tracking customer behavior, including demographics, purchase history, and participation in loyalty programs, were created. These features were crucial for understanding and segmenting customers, enabling targeted strategies and personalized marketing approaches.  5. Geographical Distribution Features: Features related to the geographical distribution of stores across California (CA), Texas (TX), and Wisconsin (WI) were generated. These features provided insights into regional trends and allowed for performance analysis based on geographic factors.  6. Marketing and Promotions Features: Features that tracked the impact of marketing campaigns and promotional events on sales and customer engagement were generated. These features were vital for assessing the effectiveness of marketing efforts and evaluating the return on investment (ROI) for promotional activities.  7. Seasonal Sales Patterns Features: Features highlighting sales patterns during different seasons and the influence of holidays and special events on store performance were created. These features were essential for understanding the seasonality of sales and optimizing strategies for different times of the year.  Rationale:  The rationale behind feature generation was to prepare the data for comprehensive analysis and modeling. By creating these features, the data became more informative and context-rich, facilitating a better understanding of store performance, customer behavior, and sales trends. These features were pivotal for subsequent analyses, predictive modeling, and evidence-based decision-making.  No specific information was provided about feature removal in the given data preparation process. However, in practice, features with high percentages of missing values or those that demonstrate little impact on the target variable may be candidates for removal. The decision to retain or remove features should be based on their contribution to the modeling and analysis objectives.  In future experiments, the generated features will serve as a solid foundation. Depending on the specific goals of those experiments, additional features or more advanced feature engineering techniques may be considered to enhance model performance and gain deeper insights. |
| **2.c. Modelling** | Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments  The experiment involved training predictive and forecasting models for sales revenue. Two primary models were considered:  1. Predictive Model (Machine Learning):  - Model Type: LightGBM, a gradient boosting framework.  - Hyperparameters Tuned:  - `n\_estimators`: The number of boosting rounds, tested with a range of values.  - `learning\_rate`: The rate at which the model adapts to errors, tested with different values.  - `max\_depth`: The maximum depth of the tree, tuned with various values.  - Rationale for Model Choice:  LightGBM is well-suited for regression tasks and can handle large datasets efficiently. It was selected for its speed, accuracy, and ability to capture complex relationships in the data.  2. Forecasting Model (Time Series Analysis):  - Model Type: SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors).  - Hyperparameters Tuned:  - Order of Autoregressive (p), Integrated (d), and Moving Average (q) components.  - Seasonal Order (P, D, Q, S).  - Rationale for Model Choice:  SARIMAX is a well-established model for time series forecasting. It was chosen to capture seasonal patterns and trends in the data.  Hyperparameter Tuning Rationale:  - The hyperparameters were tuned to find the optimal values that would yield the best model performance in terms of accuracy and predictive power.  - The number of boosting rounds (`n\_estimators`) for LightGBM, learning rate (`learning\_rate`), and maximum tree depth (`max\_depth`) were tuned to find the right balance between model complexity and predictive performance.  - For SARIMAX, the order of autoregressive, integrated, and moving average components, along with seasonal order, were tuned to capture the most significant patterns in the time series data.  Models Not Trained and Reasoning:  - Linear Regression: Linear models may not capture the complex relationships in the data as effectively as LightGBM, and they were deemed less suitable for time series forecasting.  - Neural Networks: Deep learning models were not explored due to the complexity of the problem and the availability of simpler, effective models.  Important for Future Experiments:  In future experiments, exploring more advanced ensemble methods and deep learning models, such as LSTM or Prophet for time series forecasting, could be valuable. Additionally, further hyperparameter tuning or the adoption of Bayesian optimization techniques may enhance model performance. The choice of models should align with the specific objectives and nature of the data in future experiments. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | |
| **3.a. Technical Performance** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.  The relevant performance metric for evaluating the predictive and forecasting models is the Root Mean Square Error (RMSE). Here are the RMSE values for the models:  1. RMSE using LGBMRegressor: 34.22  2. RMSE using ARIMA: 19,587.45  Analysis:  1. RMSE using LGBMRegressor: 34.22  - The RMSE of 34.22 for the LGBMRegressor model indicates that it is performing well in terms of prediction accuracy.  - The model has relatively low error, suggesting that it can effectively predict sales revenue for the given item in a specific store on a given date.  2. RMSE using ARIMA: 19,587.45  - The RMSE of 19,587.45 for the ARIMA forecasting model is considerably higher compared to the LGBMRegressor.  - The high RMSE value indicates that the ARIMA model's forecasting performance is not as accurate as desired. There is a significant discrepancy between the forecasted and actual sales figures.  Main Underperforming Cases/Observations and Potential Root Causes:  - ARIMA Model Underperformance: The ARIMA model is underperforming, as indicated by the high RMSE. Several potential root causes for this underperformance can be explored:  - Complex Seasonality: The ARIMA model might struggle to capture the complex seasonal patterns and trends in the sales data, leading to inaccurate forecasts.  - Lack of Exogenous Variables:ARIMA does not incorporate exogenous variables, such as marketing events or promotions, which can influence sales. The absence of such factors in the model may lead to inaccuracies.  - Data Quality: The ARIMA model is sensitive to data quality. Any anomalies or outliers in the data may lead to erroneous forecasts.  - LGBMRegressor Model:The LGBMRegressor model has a relatively low RMSE, indicating strong predictive performance. However, further analysis is needed to identify any specific cases or observations where it may underperform.  Recommendations:  1. ARIMA Model Improvement: To address the high RMSE in the ARIMA model, consider the following steps:  - Investigate and model more complex seasonality patterns, if present in the data.  - Include relevant exogenous variables, such as marketing events and promotions, to enhance forecasting accuracy.  - Conduct a thorough data quality assessment to identify and address anomalies or outliers.  2. LGBMRegressor Model: While the LGBMRegressor model performs well in general, it is advisable to conduct in-depth analysis on specific cases or observations where it may underperform. This analysis can help uncover any unique factors contributing to prediction errors.  3. Future Experiments: In future experiments, consider exploring alternative time series forecasting models, such as Prophet or more advanced deep learning models, to improve forecasting accuracy. Additionally, the use of Bayesian optimization for hyperparameter tuning may enhance model performance. |
| **3.b. Business Impact** | Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)  The results of the experiments have significant implications for the business objectives set earlier, which primarily revolve around sales prediction and forecasting for the retail stores in California (CA), Texas (TX), and Wisconsin (WI). Let's interpret the results and estimate the impacts of incorrect results for the business:  **Impact of Correct Results:**   1. *Accurate Sales Prediction:* Correct results are essential for accurately predicting sales for specific items in each store on a given date. This precision is crucial for optimizing inventory management, staffing allocation, and overall store performance. With accurate predictions, the business can ensure that the right products are in stock, avoiding overstocking or understocking issues. 2. *Effective Inventory Management:* Accurate forecasts enable the business to manage inventory efficiently. This means that products are available when and where they are needed, reducing the risk of stockouts and excess inventory. This efficiency can lead to cost savings and improved customer satisfaction. 3. *Enhanced Customer Satisfaction:* Accurate predictions support customer satisfaction by ensuring that customers can find and purchase the products they desire. It minimizes the chances of customers leaving empty-handed or choosing alternatives due to unavailability. 4. *Optimized Marketing Strategies:* With precise sales forecasts, the business can develop and execute marketing strategies that align with actual demand. This leads to more effective campaigns and a better return on investment (ROI).   **Impact of Incorrect Results:**   1. *Inventory Challenges:* Inaccurate sales predictions can result in inventory challenges, with a risk of overstocking or understocking. Overstocking ties up capital, while understocking can lead to lost sales and dissatisfied customers. 2. *Poor Customer Experience:* Inaccurate predictions can result in product unavailability, which can frustrate customers who cannot find what they want. This can harm customer loyalty and lead to negative reviews and word-of-mouth. 3. *Wasted Resources:* Incorrect results can lead to misallocation of staff and resources. For instance, overstaffing a store with lower sales can result in unnecessary labor costs. 4. *Marketing Inefficiency:* Poor sales forecasts can lead to inefficient marketing spending. The business may invest in campaigns that do not align with actual demand, leading to wasted resources and a lower ROI. 5. *Financial Loss:* The cumulative impact of incorrect results, including excess inventory, lost sales, and inefficient resource allocation, can result in significant financial losses for the business.   Overall, the business objective of sales prediction and forecasting is critical for efficient operations and customer satisfaction. Accurate results positively impact inventory management, customer experience, and resource allocation. Conversely, incorrect results can lead to various challenges, impacting the bottom line and customer loyalty. It is imperative for the business to continually improve its predictive models to minimize the risks associated with incorrect forecasts.  Top of Form |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.  During the experiments, several challenges were encountered and addressed. These included issues related to data quality and missing values, complex seasonality in time series data, and the absence of exogenous variables in the ARIMA model. These challenges were mitigated through data cleaning, downcasting, and hyperparameter tuning. While these issues were resolved to some extent, future experiments should focus on more robust methods for handling missing data, advanced modeling techniques for complex seasonality, and the incorporation of relevant exogenous factors to improve forecasting accuracy.  Additionally, in-depth analysis of model underperformance and addressing the complexity of seasonal sales patterns will be essential in future experiments. Furthermore, as data volume and model selection become more intricate, efficient data handling and the exploration of alternative models and hyperparameter tuning techniques are crucial for more accurate results. These challenges and solutions provide valuable insights for future experiments, ensuring the continuous improvement of predictive and forecasting models for the retail business. |

|  |  |
| --- | --- |
| 1. **FUTURE EXPERIMENT** | |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | |
| **4.a. Key Learning** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.  The experiment yielded valuable insights into the development of predictive and forecasting models for the retail business. Notably, it showcased the potential of machine learning models, particularly the LGBMRegressor, in accurately predicting sales revenue. However, it also shed light on the challenges of traditional time series analysis, emphasizing the need for advanced modeling techniques. The importance of data quality and the benefits of incorporating exogenous variables were highlighted, paving the way for more sophisticated forecasting models.  Given these insights, it is prudent to continue experimenting with the current approach. The experiment's findings provide a strong foundation for future exploration, including the adoption of advanced time series models, in-depth analysis of model underperformance, and efficient data handling for larger datasets. This approach represents a path toward continuous improvement in predictive and forecasting models for the retail business, rather than a dead end. |
| **4.b. Suggestions / Recommendations** | Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production.  The next steps and experiments for this project encompass advanced time series models, the integration of additional exogenous variables, hyperparameter tuning, ensemble modeling, deep learning models, A/B testing for marketing campaigns, and inventory optimization. These steps are ranked based on their expected uplift, with a focus on improving forecasting accuracy, optimizing inventory management, and enhancing marketing strategies. If an experiment achieves the desired outcomes, deploying the solution on Heroku would involve model export, API integration, scalability testing, continuous monitoring, documentation, user training, a feedback loop, and regular model updates to ensure accurate and efficient sales forecasting and support for business operations. This iterative approach aims to continually enhance predictive and forecasting models for the retail business. |