**Comparative Analysis of Machine Learning Models on Auto Sales Data**

|  |  |  |
| --- | --- | --- |
| Sadia Hossain  ID: 19-41347-3 *Dept Name:CSE* *Institute Name:American International University Bangladesh* Dhaka, Bangladesh sadiahossain13145@gmail.com  01302438860 | Abrar Ibne Ahsan  ID: 19-41389-3 *Dept Name:CSE* *Institute Name: American International University Bangladesh* Dhaka, Bangladesh ibneabrar.1@gmail.com  01960219292 | Sayem Sarker  ID:20-42283-1  *Dept Name:CSE* *Institute Name: American International University Bangladesh* Dhaka, Bangladesh sayemsarker43.ss@gmail.com  01708340776 |

**Introduction:**

In the ever-evolving landscape of machine learning, the ability to choose the most effective model for a specific task is crucial. This project delves into the realm of Auto Sales Data, employing various machine learning models to discern the optimal approach for predicting sales transactions. With a dataset comprising 2747 entries and 20 distinct variables, our journey involves not only implementing different models but also rigorously evaluating and comparing their performances. This report outlines the motivation, objectives, and results of the project, shedding light on the intricacies of each model's efficacy.

**Motivation/Purpose of the Project:**

The automotive industry, like many others, is increasingly leveraging data-driven insights to enhance decision-making processes. The motivation behind this project lies in the pursuit of identifying the most suitable machine learning model for predicting auto sales transactions. A robust model not only aids in forecasting sales but also contributes to optimizing inventory management, streamlining customer interactions, and ultimately boosting overall business performance.

**Dataset description:**

The dataset we choose for this project is Auto Sales Data. This dataset contains 2747 entries with 20 different variables. All of the columns have non-null values, hence there is no missing value found. The key columns include sales transaction details, customer information, product details, order status and recency information. Data types across columns, including floats, datetime and object.

**Objective:**

The primary objective of this project is to implement and compare diverse machine learning models using the Auto Sales Data. By doing so, we aim to assess the models' predictive capabilities and identify the one that exhibits superior performance in terms of accuracy, precision, and overall fit. Additionally, the project seeks to provide valuable insights into the nuances of each model's strengths and limitations when applied to real-world automotive sales data.

**Methodology:**

To achieve the project's objectives, several machine learning models were employed, each with its unique approach to predicting auto sales transactions. The models used in this project include K-Nearest Neighbors (KNN), Gradient Boosting, Linear Regression, Support Vector Machine (SVM), and Random Forest.

**Results:**

Now we check the results obtained from each model:

1. **K-Nearest Neighbors (KNN):**

**KNN** demonstrates exceptional performance with low MAE, MSE, and RMSE, indicating its suitability for predicting auto sales transactions.

* + MAE: 100.67
  + MSE: 29851.39
  + RMSE: 172.78
  + R2: 0.9902

1. **Gradient Boosting:**

**Gradient Boosting** outperforms other models with the lowest MAE, MSE, and RMSE, coupled with the highest R2 score, signifying superior predictive capabilities.

* + MAE: 113.52
  + MSE: 23458.25
  + RMSE: 153.16
  + R2: 0.9923

1. **Linear Regression:**

**Linear Regression** and **SVM** both offer reliable results, although they fall slightly behind KNN and Gradient Boosting.

* + MAE: 274.78
  + MSE: 168934.08
  + RMSE: 411.02
  + R2: 0.9446

1. **Support Vector Machine (SVM):**

Same as Linear Regression but outperformed by Gradient Boosting and KNN

* + MAE: 171.79
  + MSE: 66670.88
  + RMSE: 258.21
  + R2: 0.9781

1. **Random Forest:**

**Random Forest** exhibits potential overfitting as suggested by the notable difference between RMSE on training and testing data. Further testing and model evaluation requires to test the dataset accordingly.

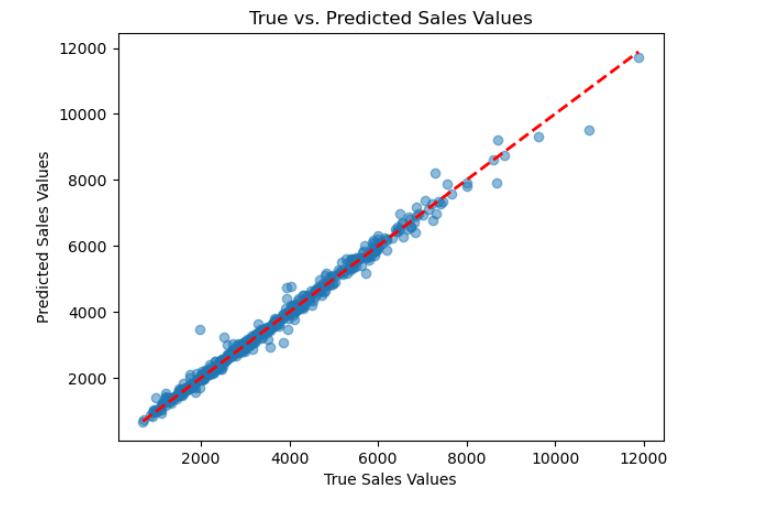
* + RMSE on training data: 66.89
  + RMSE on testing data: 281.69

**Result table:**

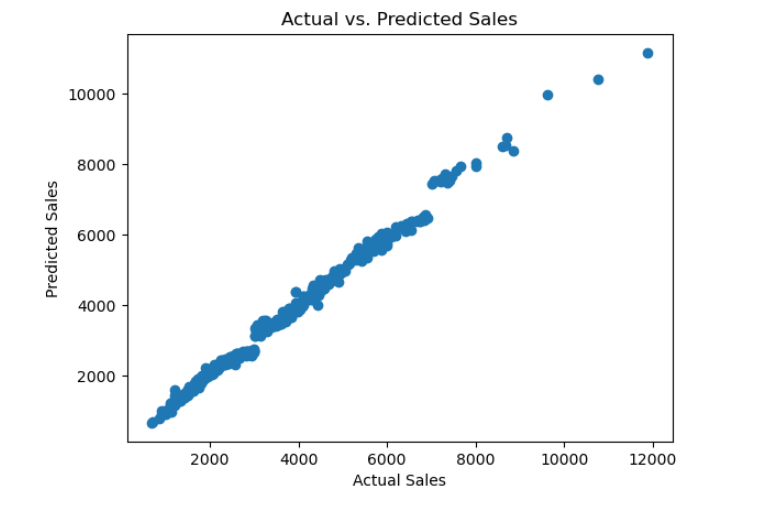
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **K-Nearest Neighbors** | **Gradient Boosting** | **Linear Regression** | **Support Vector Machine** | **Random Forest** |
| MAE: 100.67 | MAE: 113.52 | MAE: 274.78 | MAE: 171.79 | RMSE on training data:  66.89 |
| MSE: 29851.39 | MSE: 23458.25 | MSE:  168934.08 | MSE: 66670.88 | RMSE on testing data:  281.69 |
| RMSE: 172.78 | RMSE: 153.16 | RMSE:  411.02 | RMSE: 258.21 |  |
| R2: 0.9902 | R2: 0.9923 | R2: 0.9446 | R2: 0.9781 |  |

**Graph and Plots:**

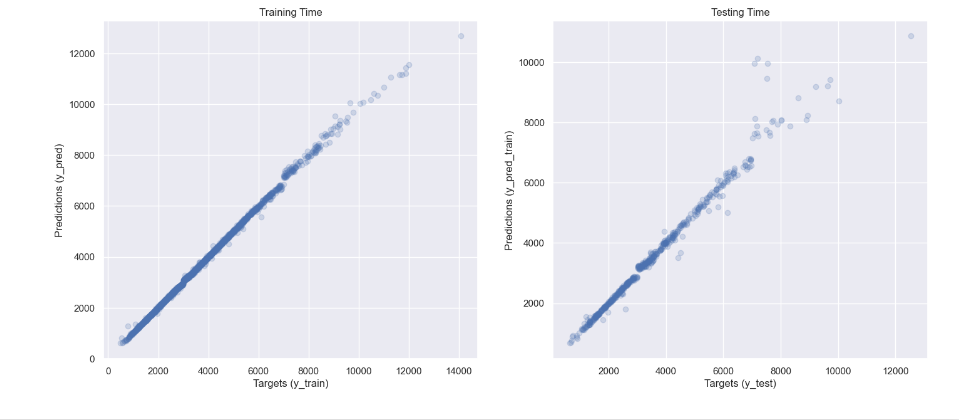
1. **KNN true vs Predicted Sales values**

****

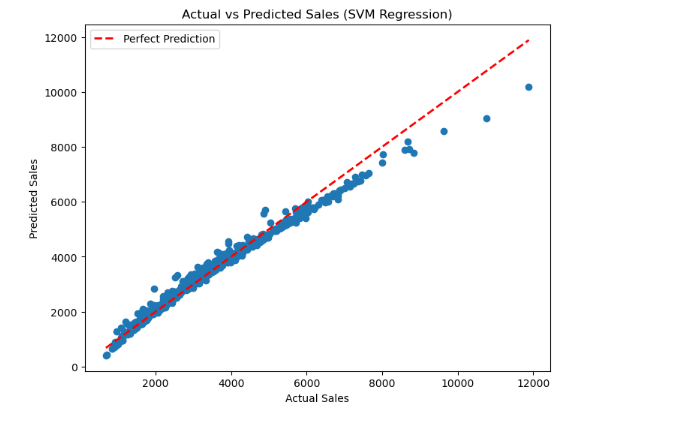
1. **Gradient Boosting Actual vs Predicted Sales**



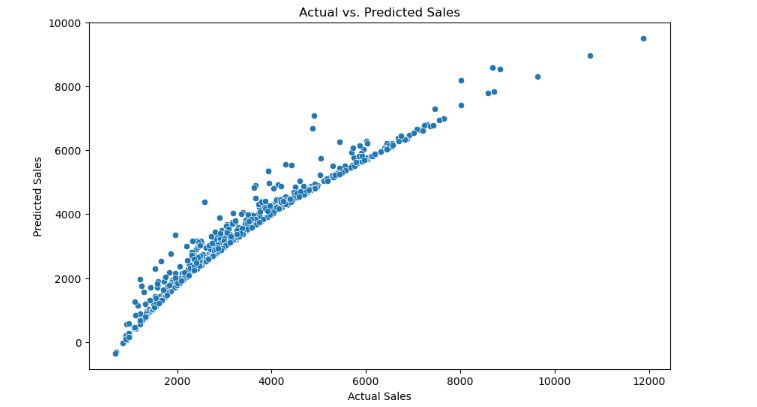
1. **Random Forest Graph**

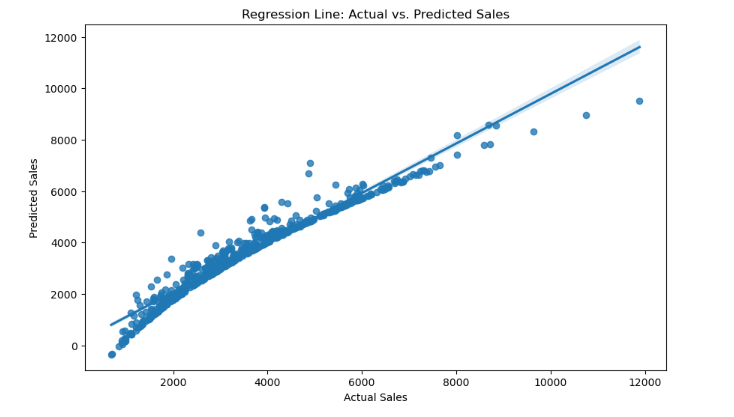


1. **Support Vector machine Actual vs Predicted Sales**

****

1. **Linear Regression**

****

****

**Conclusion:**

In conclusion, the implementation of various machine learning models on Auto Sales Data has provided valuable insights into their respective performances. **Gradient Boosting** emerges as the top-performing model, showcasing superior accuracy and precision in predicting sales transactions. **K-Nearest Neighbors (KNN)** also stands out with exceptional results, particularly in minimizing errors.

These findings are crucial for decision-makers in the automotive industry looking to optimize sales predictions and streamline business operations. As the industry continues to evolve, the ability to harness the power of machine learning models becomes increasingly indispensable.

This project not only fulfills its primary objective of identifying the best-performing model but also contributes to the broader discourse on the application of machine learning in the automotive sector. Further refinements and model tuning may be explored to enhance predictive capabilities and address specific nuances within the Auto Sales Data.

In conclusion, the pursuit of the most effective machine learning model for auto sales predictions is a dynamic and ongoing process, and this project serves as a foundational step towards achieving greater accuracy and efficiency in this domain.