Project 1:Building & Evaluating ML Algorithms

Ravi Teja Anumula (UFID: 73522006)
Dept. of Electrical and Computer Engineering
University of Florida
Gainesville, USA

***Abstract*—This project focuses on building and evaluating machine learning (ML) algorithms using a supermarket sales dataset. The dataset includes records from three branches of a supermarket, capturing customer demographics, product lines, and transaction details such as unit price, payment method, and gross income.**

Use of advanced modeling techniques, such as hyperparameter tuning and cross-validation, ensures that the predictions are both accurate and reliable. By understanding the interactions between different sales variables, businesses can tailor their strategies to optimize revenue and enhance customer loyalty.

For gross income prediction, multiple linear regression models are employed to analyze factors such as unit price, quantity, time of purchase, and product lines. Lasso regularization is applied to highlight the most impactful drivers of revenue, helping businesses identify key opportunities for increasing profitability. Additionally, customer behavior is examined through classification models, which explore how gender, product preferences, and payment methods influence purchasing patterns. A detailed analysis of Branch C specifically uncovers insights into male customers' behaviors, offering opportunities for more personalized promotions and targeted marketing.

***Index Terms*—Prediction, salesmarketCSV, Regression and classification, Hyperparameter tuning, cross validation**

I. Introduction

The project's scope revolves around maximizing profitability and operational efficiency by understanding customer behavior, product performance, and sales trends through data-driven insights. The dataset provided includes detailed sales records such as customer demographics, product lines, and transactional data, all of which are valuable for uncovering patterns that can drive business decisions.

The primary objectives of the project are:

A. Revenue Optimization via Predictive Analytics

By predicting gross income based on variables like unit price, quantity sold, product line, and customer demographics, the business can anticipate revenue fluctuations and identify high-performing products. The predictive models help guide pricing strategies, enabling the business to adjust prices dynamically to maximize gross income.

B. Customer Segmentation and Targeted Marketing:

Customer classification based on attributes such as gender, product preferences, and payment methods allows for precise segmentation. The insights can enable personalized marketing campaigns, helping the business to engage customers more effectively. For example, understanding which product lines are preferred by different customer demographics can help tailor promotions and discounts, leading to higher conversion rates.

C. Optimizing Product Inventory and Sales Forecasting

By analyzing product line performance, management can optimize inventory levels and reduce stockouts or overstocking. Sales forecasting models allow the business to anticipate high-demand periods and make data-backed decisions about which products to promote or replenish.

D. Branch-Specific Sales Strategies

The analysis of Branch C, in particular, provides insights into customer behavior and product preferences specific to that location. These insights allow branch managers to implement strategies that are tailored to the unique demands of their customer base, optimizing sales and enhancing customer satisfaction.

II. Methodology

The project methodology is structured around utilizing data preprocessing, model selection, and performance evaluation techniques to derive insights, with a focus on business outcomes.

A. Data Preprocessing

a) Feature Engineering: Business-relevant features such as the time of purchase and the day of the week were engineered from the dataset. The date was converted into the day of purchase (e.g., Monday, Tuesday), and the time was grouped into slots (morning, afternoon, evening). These new features allow for a granular understanding of when customers are most likely to purchase, enabling the business to schedule promotions or adjust staff levels accordingly.

b) Categorical Encoding: Customer demographics (such as gender and payment method) and product attributes (like product line) were encoded into a format suitable for machine learning models. One-hot encoding was used to represent these categorical features, preserving the business context of each attribute (e.g., whether payment method influences product choices).

B. Model Selection and Hyperparameter Tuning

a) Regression Models for Revenue Prediction: Linear Regression and Lasso Regression models were employed to predict gross income and unit prices. Lasso regularization was particularly useful for eliminating irrelevant features, ensuring that the model only includes business-critical variables.

This reduced complexity provides clearer insights into which factors directly drive revenue, helping management focus on high-impact areas.

b) Classification Models for Customer Segmentation: Logistic Regression, Decision Tree and Random Forest classifiers were utilized to categorize customers based on their demographics and behavior, such as customer type and gender along with their day of purchase. Interaction features of degree two were used to explore how combinations of attributes (e.g., payment method and product line) influence customer behavior. This helps the business understand how multiple factors combine to affect buying decisions, enabling multi-faceted marketing strategies.

C. Evaluation Metrics and Business Implications

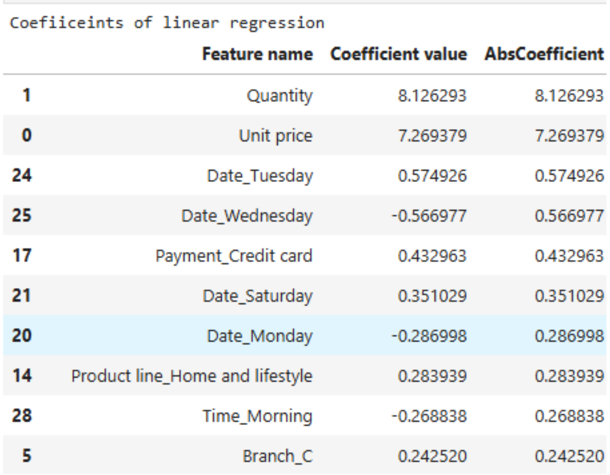
a) Performance Metrics: For regression models, the R² score and 95% confidence interval were used to evaluate how accurately the models predict gross income, giving the business an understanding of how well their revenue forecasting model performs under different conditions. For classification models, accuracy and its 95% confidence interval were used to measure how well customer segments are identified, allowing the business to trust these models for personalized marketing strategies.

b) Interpretability: Lasso Regression and Decision Trees were chosen for their interpretability. These models provide insights into which features are most influential, guiding business decisions on customer targeting and product offerings.

III. Observations

The regression models provide valuable insights into how different factors affect gross income and unit price.

A. Multiple linear regression without lasso to predict gross income



```
Coefiiceints of linear regression
```

	Feature name	Coefficient value	AbsCoefficient
1	Quantity	8.126293	8.126293
0	Unit price	7.269379	7.269379
24	Date_Tuesday	0.574926	0.574926
25	Date_Wednesday	-0.566977	0.566977
17	Payment_Credit card	0.432963	0.432963
21	Date_Saturday	0.351029	0.351029
20	Date_Monday	-0.286998	0.286998
14	Product line_Home and lifestyle	0.283939	0.283939
28	Time_Morning	-0.268838	0.268838
5	Branch_C	0.242520	0.242520

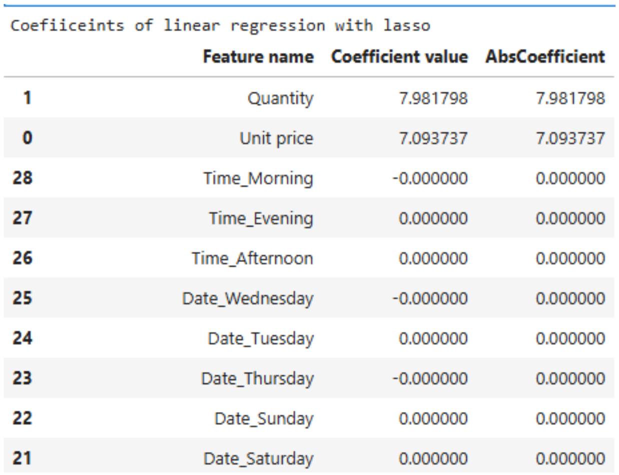
Fig. 1. Weights of Features for predicting gross income without lasso

In the linear regression model (without Lasso regularization), the analysis reveals that gross income is primarily driven by two key factors: **Quantity and Unit Price**. As either of these variables increases, gross income rises proportionally, highlighting that optimizing pricing strategies and encouraging higher purchase quantities can directly boost revenue.

Further insights show that specific attributes like Day of Purchase, Product Line, and Time Slot significantly impact gross income as well. Notably:

- **Tuesday** is the most profitable day of the week, indicating that targeted promotions or marketing efforts on this day could amplify sales.
- **The Home and Lifestyle** product line generates the highest gross income, making it a strategic focus for upselling and cross-selling opportunities.
- **Morning time slots** also contribute notably to gross income, suggesting that customer engagement and marketing campaigns timed in the morning can maximize revenue potential.

B. Multiple linear regression with lasso to predict gross income



```
Coefiiceints of linear regression with lasso
```

	Feature name	Coefficient value	AbsCoefficient
1	Quantity	7.981798	7.981798
0	Unit price	7.093737	7.093737
28	Time_Morning	-0.000000	0.000000
27	Time_Evening	0.000000	0.000000
26	Time_Afternoon	0.000000	0.000000
25	Date_Wednesday	-0.000000	0.000000
24	Date_Tuesday	0.000000	0.000000
23	Date_Thursday	-0.000000	0.000000
22	Date_Sunday	0.000000	0.000000
21	Date_Saturday	0.000000	0.000000

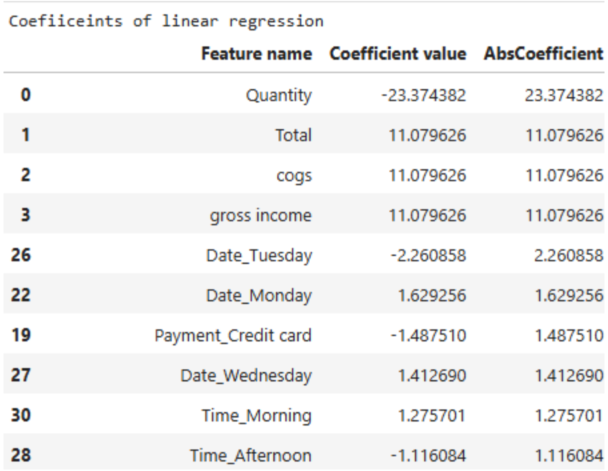
Fig. 2. Weights of Features for predicting gross income with lasso

In the linear regression model with Lasso regularization, the analysis simplifies the factors influencing gross income. The Lasso regularization removes the impact of all other variables, highlighting that **Quantity and Unit Price** are the sole significant drivers of gross income.

This means that, from a business perspective, focusing on strategies that directly influence these two factors—such as optimizing product pricing and encouraging customers to purchase higher quantities—will have the most substantial effect on increasing revenue

C. Multiple linear regression without lasso to predict unit price

The weights of features for linear regression without lasso suggests that Unit Price tends to be closely tied to the **quantity** sold, as well as **overall revenue and costs**, indicating that increasing volume or optimizing pricing strategies can lead to better pricing decisions. Additionally, focusing on gross



Coefiiceints of linear regression

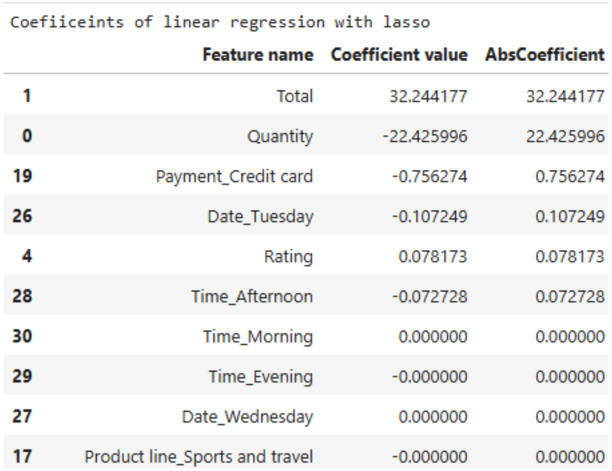
	Feature name	Coefficient value	AbsCoefficient
0	Quantity	-23.374382	23.374382
1	Total	11.079626	11.079626
2	cogs	11.079626	11.079626
3	gross income	11.079626	11.079626
26	Date_Tuesday	-2.260858	2.260858
22	Date_Monday	1.629256	1.629256
19	Payment_Credit card	-1.487510	1.487510
27	Date_Wednesday	1.412690	1.412690
30	Time_Morning	1.275701	1.275701
28	Time_Afternoon	-1.116084	1.116084

Fig. 3. Weights of Features for predicting Unit price without lasso

income and COGS helps refine profit margins and pricing strategies.

Unit Price is also influenced by **Tuesday, Morning time slots, and the Fashion Accessories product line**. This indicates that businesses could benefit from targeting specific days, times, or product categories where unit prices are more responsive, helping to optimize sales pricing strategies

D. Multiple linear regression with lasso to predict unit price



Coefiiceints of linear regression with lasso

	Feature name	Coefficient value	AbsCoefficient
1	Total	32.244177	32.244177
0	Quantity	-22.425996	22.425996
19	Payment_Credit card	-0.756274	0.756274
26	Date_Tuesday	-0.107249	0.107249
4	Rating	0.078173	0.078173
28	Time_Afternoon	-0.072728	0.072728
30	Time_Morning	0.000000	0.000000
29	Time_Evening	-0.000000	0.000000
27	Date_Wednesday	0.000000	0.000000
17	Product line_Sports and travel	-0.000000	0.000000

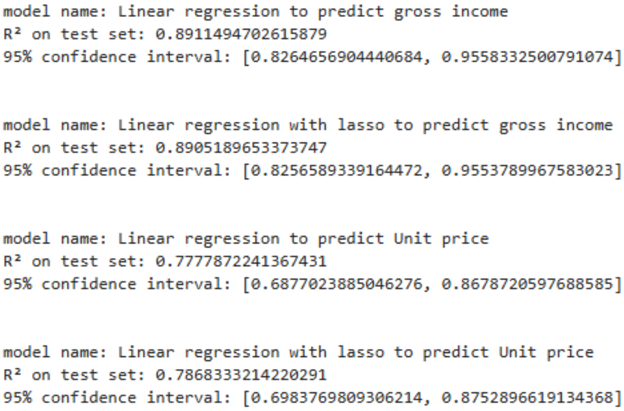
Fig. 4. Weights of Features for predicting Unit price with lasso

In the linear regression model with Lasso regularization, Unit Price is primarily influenced by **Total sales**, followed by **Quantity** sold. The use of Lasso regularization significantly reduces the impact of most other features, highlighting a more focused relationship between Unit Price, Total sales, and Quantity.

From a business perspective, this suggests that the Total sales and Quantity are the key drivers of Unit Price, and optimizing these factors could lead to better pricing strategies. By focusing on higher sales volumes or adjusting quantities, businesses can more effectively set prices.

Additionally, the influence of specific factors like **Tuesday** and **Afternoon** time slots remains, though their impact is relatively small. This indicates that while day and time do have some effect on Unit Price, their contribution is not as significant as the core drivers such as sales totals and quantities. Businesses can still consider adjusting strategies for specific days or times to capture small pricing opportunities, but the primary focus should remain on optimizing Total sales and Quantity for more substantial impact.

E. Performance measurement of regression models



```
model name: Linear regression to predict gross income
R² on test set: 0.8911494702615879
95% confidence interval: [0.8264656904440684, 0.9558332500791074]


model name: Linear regression with lasso to predict gross income
R² on test set: 0.8905189653373747
95% confidence interval: [0.8256589339164472, 0.9553789967583023]


model name: Linear regression to predict Unit price
R² on test set: 0.7777872241367431
95% confidence interval: [0.6877023885046276, 0.8678720597688585]


model name: Linear regression with lasso to predict Unit price
R² on test set: 0.7868333214220291
95% confidence interval: [0.6983769809306214, 0.8752896619134368]
```

Fig. 5. R2 score and confidence interval for Regressions

a) Linear Regression without Lasso for Gross Income:

- **R² = 0.89** (**Confidence Interval: [0.83, 0.96]**)
- High accuracy, explaining 89% of gross income variation. Reliable for financial forecasting, revenue prediction, and budget planning.

b) Linear Regression with Lasso for Gross Income:

- **R² = 0.89** (**Confidence Interval: [0.83, 0.96]**)
- Similar performance to non-regularized model, with enhanced stability and improved interpretability, ideal for reducing overfitting in predictive applications.

c) Linear Regression without Lasso for Unit price:

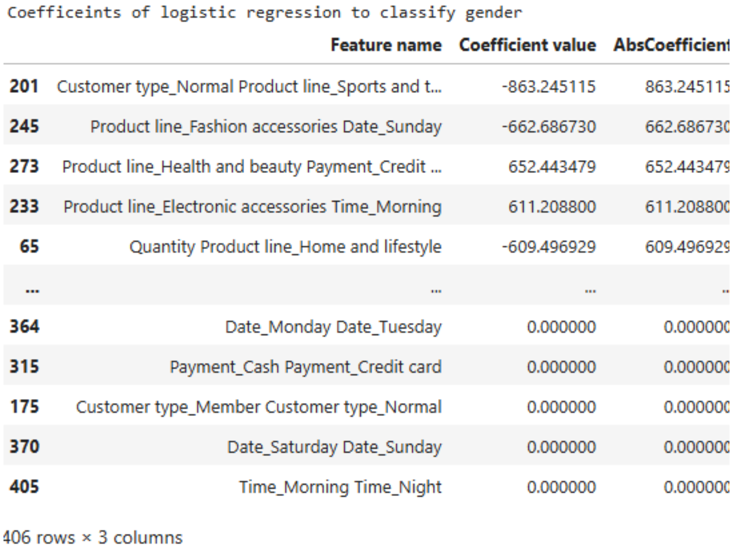
- **R² = 0.78** (**Confidence Interval: [0.69, 0.87]**)
- Moderate accuracy, capturing 78% of unit price variation. Suitable for basic pricing optimization, but further model refinement is needed for dynamic pricing strategies.

d) Linear Regression with Lasso for Unit Price::

- **R² = 0.79** (**Confidence Interval: [0.70, 0.88]**)
- Slight improvement over non-regularized model. Effective for streamlined pricing decisions with a focus on the most influential factors, boosting profit margin management.

F. Logistic regression to predict gender classification

The analysis reveals that interactions between multiple features significantly influence the model's ability to distinguish between genders. By considering not only individual features but also their combined effects, the model identifies



```
Coefficeints of logistic regression to classify gender
```

	Feature name	Coefficient value	AbsCoefficient
201	Customer type_Normal Product line_Sports and t...	-863.245115	863.245115
245	Product line_Fashion accessories Date_Sunday	-662.686730	662.686730
273	Product line_Health and beauty Payment_Credit ...	652.443479	652.443479
233	Product line_Electronic accessories Time_Morning	611.208800	611.208800
65	Quantity Product line_Home and lifestyle	-609.496929	609.496929
...	...	...	...
364	Date_Monday Date_Tuesday	0.000000	0.000000
315	Payment_Cash Payment_Credit card	0.000000	0.000000
175	Customer type_Member Customer type_Normal	0.000000	0.000000
370	Date_Saturday Date_Sunday	0.000000	0.000000
405	Time_Morning Time_Night	0.000000	0.000000

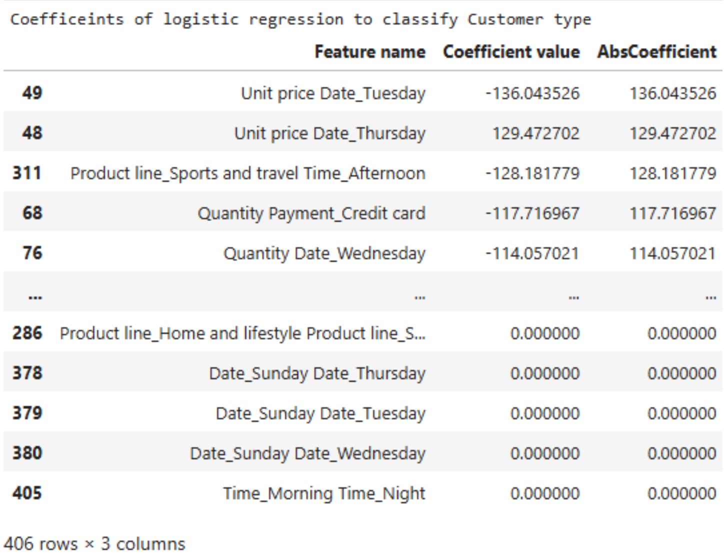
406 rows × 3 columns

Fig. 6. parameters values for all attributes (and its 2nd-order interactions)

more complex patterns that are indicative of gender-specific behaviors.

From a business perspective, this approach uncovers valuable insights into how different factors, such as product line, payment method, and purchase day/time, interact to influence gender-based purchasing decisions. For instance, interactions like product line and payment method may highlight that male customers are more likely to make purchases in certain categories or use specific payment methods at particular times of the day.

G. Logistic regression to predict Customer type classification



```
Coefficeints of logistic regression to classify Customer type
```

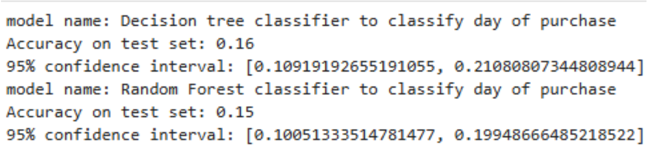
	Feature name	Coefficient value	AbsCoefficient
49	Unit price Date_Tuesday	-136.043526	136.043526
48	Unit price Date_Thursday	129.472702	129.472702
311	Product line_Sports and travel Time_Afternoon	-128.181779	128.181779
68	Quantity Payment_Credit card	-117.716967	117.716967
76	Quantity Date_Wednesday	-114.057021	114.057021
...	...	...	...
286	Product line_Home and lifestyle Product line_S...	0.000000	0.000000
378	Date_Sunday Date_Thursday	0.000000	0.000000
379	Date_Sunday Date_Tuesday	0.000000	0.000000
380	Date_Sunday Date_Wednesday	0.000000	0.000000
405	Time_Morning Time_Night	0.000000	0.000000

406 rows × 3 columns

Fig. 7. parameters values for all attributes (and its 2nd-order interactions)

From observations, it shows that these 2nd order interactions can drive highly targeted marketing and customer retention strategies. For instance, the model show that certain combinations of purchase day and unit price strongly correlate with a customer being a normal buyer. These insights enable businesses to craft personalized offers, promotions, or loyalty programs that specifically appeal to these high-value segments, potentially boosting customer memberships and lifetime value.

H. Classifier to predict day of purchase



```
model name: Decision tree classifier to classify day of purchase
Accuracy on test set: 0.16
95% confidence interval: [0.10919192655191055, 0.21080807344808944]
model name: Random Forest classifier to classify day of purchase
Accuracy on test set: 0.15
95% confidence interval: [0.10051333514781477, 0.19948666485218522]
```

Fig. 8. Accuracy and its confidence interval for predicting day of purchase)

The Decision Tree and Random Forest classifiers were used to predict the day of purchase based on customer data. Both models achieved relatively low performance, with accuracy values of 16% for the Decision Tree and 15% for the Random Forest model. These figures are far below what would be considered useful for actionable business insights, as they suggest the models are struggling to accurately predict the day of the week when a customer is most likely to make a purchase.

From a business perspective, this low accuracy implies that predicting the day of purchase is not easily discernible from the available features in the dataset. This may indicate that factors like customer behavior, transaction history, or seasonality—which could have significant predictive power—are not being sufficiently captured or modeled. The inability of both classifiers to accurately predict purchase days suggests that there may not be strong correlations between the features in the current dataset and the day a purchase occurs.

IV. Conclusion

The project successfully developed predictive models for gross income and unit price with high accuracy, explaining 89% and 78-79% of the variance, respectively. Lasso regularization improved interpretability, highlighting key drivers like quantity and unit price. However, models for day of purchase prediction using Decision Tree and Random Forest showed lower accuracy (15-16%), indicating potential for refinement. Overall, the models provide valuable insights for financial forecasting and pricing optimization, supporting data-driven decisions that can enhance profitability and operational efficiency, with room for further improvement in customer behavior predictions.