Statistical Project

Luis Marcos López Casines
Laia Porcar Guillamón
Philippe Robert

University of Padova
Department of Mathematics

Table of contents

1. Introduction and motivation
2. Data Cleaning Manipulation
3. Relevant Questions
4. Exploratory Data Analysis (EDA)
5. Linear Regression
6. Multiple linear Regression
7. Classification
8. Conclusion

Introduction

A common task for data scientist is to analyze sales and profits, as well as the different features that are attached to them, of an entity to extract meaningful insights.

The objective of this work is to analyze the purchases made in an anonymous online store. By doing so, we aim at answering relevant questions to have a better understanding of what drives the sales and the profits.

Data Cleaning and Manipulation

- Superstore dataset
  - Taken from Kaggle, publicly available for educational purposes.
  - Rows: 9994, 21 columns.
  - 5009 order IDs, 3 categories, 17 subcategories, 532 cities, 49 states, 4 regions, 4 ship modes, 793 different customers.

- Data Cleaning
  - No missing value of any type, no duplicates.
  - Remove columns: Customer Name, Country, Postal Code.
  - Convert dates to date format.
  - Add columns: Gross Margin, Processing Time, Year, Month.
  - Outlier detection.

Outlier detection

Process

1. Summary of numerical variables.
2. Histograms of distributions.
3. Transformation of Profit and Sales.
4. Boxplots.

Log transformations

Boxplots

- Potential outliers based on the plots.
- Based on context no significant outliers.

Relevant Questions

As a starting point to develop the following sections of the work, we will try to answer the next questions.

- Which subcategories are more popular? Do they also generate more profit?
- Do different segments prefer different ship modes?
- Which segment is generating more profit? In which region?
- In which states and cities are there more orders? Do they also generate more profit?
- How do orders, sales and profit behave throughout the year and during holidays? Can we identify some patterns?
- Which is the top-10 clients according to purchases and to which segment do they belong? Are they the ones that generate more profit?
- Does the ordered quantity change when there is a discount?
- How does the profit change for different discount ranges? What would be a threshold discount for positive/negative profits?
- What is the relation between profit and sales? Is it linear?
- How accurately can we estimate this relation?
- How accurately can we predict future profits based on sales?
- What features are more relevant when determining the profit?
- Is there any interaction among the different features?
- Is it possible to predict if a purchase will result on positive or negative profit? Which features are more relevant for this task? How accurate is the prediction?

Addressing relevant questions

How are we going to address them ?

Exploratory Data Analysis

Analysis of discrete variables, top 10 clients analysis, analysis of continuous variables, correlation study.

Linear Regression

Linear regression on Profits vs Sales.

Multiple Linear Regression

Dimensionality reduction: Backward selection, Lasso regression, Ridge regression, interaction terms, nonlinear terms. Prediction on test set.

Classification

Feature selection, logistic regression, Bayes classifier , linear and quadratic discriminant analysis.

Exploratory Data Analysis (EDA)

Before making any assumption, we go through the process of exploring the data. This is the corner stone of our analysis and will pave the way to further stage of examination of our dataset.

Analysis of discrete variables

- Office Supplies represent more than half of the total orders.
- The three most popular subcategories represent 39% of the total.

Analysis of discrete variables

- The two cities with most orders (LA, NYC) belong to the two states with most orders (CA, NY).
- Consumer dominate the number of order and profit.
- Home Office generates the less losses.
- For all categories, segments and years the shipping mode's ranking is:
  1. Standard Class
  2. Second Class
  3. First Class
  4. Same Day

Analysis of discrete variables

- Each category represents around 1/3 of the sales, with a small domination of Technology.
- The two highest months in number of orders and profit are November and December.
- Positive profit dominated by Technology.
- Losses dominated by Office Supplies.
- The two worst months in number of orders and profit follow the two best ones: January and February.

Analysis of discrete variables

- Copiers generate the largest net profit.
- Tables generate the lowest net profit.

Analysis of discrete variables

- States with more orders are mostly the most populated.
- Case TX, PA and FL: big amount of orders but low profit.

Analysis of discrete variables

- Only 2015 had a (small) decrease on sales. However, its profit still increased.
- 2 holiday months represent almost half of orders of the other 10 months together.

- Furniture yields the worst profit.
- Technology and Office Supplies have similar behavior.

- Central region is the only one with negative profit in a category: Furniture.
- The West and South region have the highest median profit.
- Technology has the highest median profit.

- Quantity of Fasteners, Bookcases and Phones increase when discounts are applied.
- Quantity of Machines bought decreases when more discounts are applied.
- Furniture generates the highest losses when discounted.
- Office Supplies is the only one with discount over 75%.

Analysis of discrete variables

Top-10 w.r.t purchases.

- 6/10 are Consumer.
- Represent 6.6% of total orders.
- 5/10 generate significantly losses.

Top-10 w.r.t positive profit.

- 6/10 are Consumer.
- The most "profitable" client is Corporate.

Top-10 w.r.t negative profit.

- Client that causes more losses is Consumer and buys mostly Technology.
- 2 Home Office, 4 Consumer, 4 Corporate.

Top-10 w.r.t sales.

- Client that represents the most sales is Corporate.
- Also is one of the top-10 clients according to purchases.

Furniture is not determinant to be in the Top-10 of orders, sales nor profit.

Analysis of continuous variables

Analysis of continuous variables

Profit vs Sales

Not linear for sure...

Analysis of continuous variables

Let's see for which category we identify a linear behavior.

Correlation Analysis

Ranked Cross−Correlations

8 most relevant

Linear Regression

Focus on Profit vs Sales for copiers.

- Train-test split of 80-20%.
- $R^2 = 0.96$
- F-statistic: 3.307×10^4, p-value: $< 2.2 \times 10^{-16}$
- Expected change in Profit due to a unit change in Sales: 0.48.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-1.028	0.204	-5.027	8.74e-07
Sales	0.479	0.003	181.848	<2e-16

	2.5%	97.5%
Intercept	-1.430	-0.626
Sales	0.474	0.484

Prediction carried out on test set:

- RMSE = 8.64.
- Normalized RMSE = 2.25%

Diagnostic plots

Heteroskedasticity, no linearity and three potential outliers.

Multiple linear Regression

- Train-test split: 80-20%.
- Backward selection of features.
- Dependent variables: Profit.
- Predictors: Sales, Quantity, Segment, Discount, Processing.Time, Ship.Mode, Region, Sub.Category, Gross.Margin.
- Use of 5-fold CV.

First results:

- F-statistic: 156.8. Evidence against the null hypothesis. At least one of the selected features must be related to Profit.
- Processing.Time, Ship.Mode, Segment $\rightarrow$ small p-value
- VIF okay except for Gross.Margin and Discount $\rightarrow$ discard Gross.Margin.
- Adjusted R^2 of 0.36.

Diagnostic plots

Lasso Regression and Ridge Regression

- Confirm Processing.Time, Ship.Mode, Segment not significant for the model.
- Sales could also be discarded: but wait till next steps.

Interaction terms

- Interaction between Sales and Discount greatly improves the adjusted R^2, no multicollinearity, low p-value. Checked with ANOVA.

Nonlinear terms

- Sales up to order 3 improves the adjusted R^2, VIF okay, low p-values. Checked with ANOVA.

Final variable selection

- High p-value of Region in final model. Discard it.
- Final choice: poly(Sales, 3), Sales:Discount, Sub.Category, Gross.Margin.
- F-statistic: 1795 $\rightarrow$ 10 times larger than initial model. Almost null p-value.
- Adjusted R^2: 0.83
- Prediction on test set:
  - RMSE: 91.74
  - Normalized RMSE: 0.61%

Final diagnostic plots

Classification

Binary classification problem: predict negative/positive profit.

Unbalanced data: 18% negative class.

Feature selection

- Based on Akaike Information Criteria (AIK) and McFadden's R^2.
- Focus on Discount, Sales, Sub.Category, Processing.Time and Region.

Logistic Regression

- Always use 5-fold CV.
- Unbalanced data.
- Balanced Data: two approaches.
  - ROSE library: sampling methods and smoothed bootstrap.
  - Ovunsample: mix of under-sampling and over-sampling.

Method	AIC	McFadden's R^2	Accuracy	Precision Pos/Neg	Recall Pos/Neg	F1 score Pos/Neg	AUC
Unbalanced	1541.20	0.77	0.94	0.98/0.76	0.95/0.91	0.97/0.83	0.87
ROSE	2856.60	0.70	0.92	0.92/0.95	0.99/0.73	0.95/0.83	0.94
Ovun.sample	1960.9	0.79	0.92	0.92/0.95	0.98/0.73	0.95/0.82	0.93

Try other classification models, using unbalanced data.

Bayes Classifier

- Similar performance to logistic regression on unbalanced data.
- AUC: 0.85

Linear Discriminant Analysis

- Similar performance to logistic regression and Bayes.
- AUC: 0.87

Quadratic Discriminant Analysis

- Similar performance to all previous models.
- AUC: 0.85

Conclusions

Summary of the keys analysis conclusion.

- Among the 5 most popular subcategories, only Paper and Art see a subsequent profit increase when the sales increase.
- Furniture has high profit volatility throughout the years and generates net losses from 15% discount, Office Supplies and Technology from 30%.
- Top 10 clients who cause more losses buy mostly Technology, which in turn has the highest median profit per order.
- Optimal places to place warehouse: CA, TX, PA.
- Consumer segment dominates top 10 clients for number of orders, net positive and negative profit. Corporate dominates Sales.
- Multiple linear regression yields poly(Sales,3), Sales:Discount, Sub Category and Gross Margin as most relevant features to predict profit. Adjusted R^2 of 0.83 and normalized RMSE of 0.61%. Heteroskedasticity $\rightarrow$ can't trust the confidence intervals of estimates.
- Classification model can predict pos/neg profit with an AUC up to 94%.

Thank you for your attention.