ONLINE RETAIL – Pre-project of Machine Learning

Adrien de Mailly Nesle, Mathieu Cowan and Alaric de Bastard (DIA 3)

1. Business Challenges and State of the Art

- **Challenge**: For this dataset, the technical challenge is to be able to predict multiple elements which are particularly important. We want to predict the revenue for a certain type of product and for a certain period so we can make decisions based on it. We can use different strategies to categorize our products and make our prediction based on those categories. A good example of study could be during the Christmas period. We also would like to predict the state of the stocks because it is important for the sustainability of retail.

- **State of the Art**: In online retail, machine learning enhances customer insights and business strategies. Customer segmentation uses clustering and deep learning to group customers by purchasing behavior, while customer lifetime value (CLV) prediction employs regression and survival analysis to estimate future revenue. Recommendation systems leverage collaborative filtering and deep learning for personalized product suggestions. Anomaly detection helps identify fraud through outlier detection and unsupervised learning, and churn prediction applies classification and sequential models to forecast customer attrition. Demand forecasting uses time series and machine learning to optimize inventory, aligning stock with anticipated demand. These methods together drive better decision-making and customer engagement in retail.

2. Data Description and Sources

- **Data Description:**

- **InvoiceNo:** a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation

- **Stock:** a 5-digit integral number uniquely assigned to each distinct product

- **Description:** product name

- **Quantity:** the quantities of each product (item) per transaction

- **InvoiceDate:** the day and time when each transaction was generated

- **UnitPrice:** product price per unit

- **CustomerID:** 5-digit integral number uniquely assigned to customer

- **Country:** the name of the country where each customer resides

- **Data Sources:**

The dataset comes from the University of **UC Irvine (UCI) Machine Learning Archives**, it is the Online Retail Dataset downloaded in 2015. To be more precise; this is a transactional dataset which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

3. Business Objectives and Scope

- **Primary Objective** :

For any product, the goal is to make the most profit possible out of it and to properly manage their stocks. In this optic, there are strategies like increasing or lowering prices, creating promotions etc. And those actions are made possible by having information on the actual revenues that the products provide. We also have the ambition to determine the revenue for any product at any time and to make sure that we can have an accurate prediction by considering every parameter, the period of the year, the status of the client, etc. This would result in us implementing a strategy of customer relationship management through our models.

- **Functionality**: The model should make predictions on the revenue or on the state of the stock, especially if it is critical for the different types of products for different year periods.

- **Scope regarding each step of the project :** In this part, we analyze the complexity of the project scope step by step regarding our business problem, whether it is regarding the optimization of the revenue, or regarding the optimization of the business strategy.

Step 1 : To analyze our data, we will process in multiple subtasks:

- Data Preprocessing: First, we need to take out the rows concerning the transactions which were cancelled (with the InvoiceNo starting by "C") because they obviously do not have any positive or negative consequence on the revenue or the stocks. Then we will create a new column for the revenue where we will multiply the quantity of product in the transaction by the unit price. There will be cases where we miss essential information like the description of the product. This information is linked to a stock ID, which means that we can find the missing description of a product with a stock ID and inversely. Unfortunately, there will be cases where we will not have enough information to use the row in which case we will drop it. Also, something we could do to make a more precise database is to add new features to our dataset with the attributes of the InvoiceDate. For instance, we can add a "season" attribute or a "month" attribute, which can permit a better visualization of the analysis of the data.

- Exploratory Data Analysis: Analyze data distributions, correlations, between variables and identify any multicollinearity or strong predictors for revenue or the state of stocks.

When it comes to **model training**, we will use models such as classification and regression to predict the revenue and the quantity of sold products (modification of stocks) per product regarding the period of the year and the horary period of the day.

Step 2 : To advance further in our study, we will make sure to enhance our ability to make predictions regarding the situation of the retail to be able to answer as precisely as possible to the business problem. Therefore, we will try to use advanced models to make predictions on the better period in which we will put discounts in place for x and y products.

Step 3 : We will now use more advanced models and principles regarding Machine Learning/Deep Learning to be able to answer even more precisely to the business problem.

When it comes to Model Training, we will use embedding of the description of the products. We will use it to implement a **regression model** with the revenue as target. We can also use classification using the distribution thresholds of the revenue (e.g. 20% of data between 100 and 1000 dollars of revenue).

For Model Evaluation, we split the dataset into subsets for each country over the entire period. Then we perform cross validation on each country set and we want to minimize the RMSE. We can make the same process with time periods instead of countries.

Conclusion of the scope: Based on the Online Retail dataset, our project will use Machine Learning tools such as regression and classification models to create a tool capable of predicting revenue on an online retail platform. It will allow us to better handle the data that flows within the platform and to then make adapted business decisions based on this study.

Work plan :

Stage	Task	Start Date	End Date	Task repartition
Stage 1: Standard Solutions	**Data Analysis & Preprocessing**: Clean dataset (remove cancellations, fill missing data, create revenue column, extract date features, etc.)	08-nov	10-nov	**Mathieu Cowan**
	Exploratory Data Analysis (EDA): Correlations, multicollinearity, identify strong predictors	10-nov	12-nov	**Alaric de Bastard**
	Model Training: Apply classification and regression algorithms, and train on predictions (revenue, stock changes)	12-nov	13-nov	**Mathieu Cowan**, **Alaric de Bastard**
	Learning & Testing Plan: choose algorithms for model and methods for overfitting control (cross-validation)	13-nov	13-nov	**Adrien de Mailly Nesle**
	Evaluation: Assess model performance and conduct error analysis	13-nov	13-nov	**Adrien de Mailly Nesle**
	Documenting Results: formalisation methodology, results obtained and criticisms, conclusion and prospects	13-nov	15-nov	**All Members**
Stage 2: Improving Solutions	**Advanced Model Implementation**: Implement more advanced algorithms from class (more complex approach)	16-nov	18-nov	**Mathieu Cowan**, **Adrien de Mailly Nesle**
	Learning & Testing Plan: add columns if necessary, apply model, apply methods for overfitting control	18-nov	20-nov	**Alaric de Bastard**
	Evaluation: Measure performance improvements, compare against Stage 1	21-nov	22-nov	**All Members**
	Ensemble Learning: Implement ensemble models to improve decision-making (e.g., bagging, boosting)	22-nov	24-nov	**Mathieu Cowan**, **Alaric de Bastard**
	Documenting Results: algorithm limitation highlighting, explanations of the solution algorithm, comparison of results with previous step, conclusion on step	25-nov	29-nov	**All Members**
Stage 3: Advanced Solutions	**Algorithm Selection**: Choose and justify a more complex approach/algorithm	30-nov	02-déc	**Adrien de Mailly Nesle**
	Implementation: Implement advanced model, apply evaluation, and compare with previous results	02-déc	04-déc	**Adrien de Mailly Nesle**
	Final Report: Document project summary, methodology, results, discussion, and conclusions	05-déc	06-déc	**All Members**

5. Conclusion: By working on this project, we will surely demonstrate how machine learning can help in optimizing retail online operations, especially when it comes to revenue and stock prediction to help the decision-making. We will use different models to be able to predict revenue for specific products/periods. Thus, we will hopefully be able to have a model which can help avoid shortages or overstocks for the stock levels and will help to optimize the revenue with price-optimizing strategies.

6. References

References and potential future references include:

- The dataset: https://archive.ics.uci.edu/dataset/352/online+retail
- Market Segmentation and its Impact on Customer Lifetime Value Prediction
- Customer Lifetime Value: Methods and Models
- Customer Churn Prediction in E-commerce: A Survey