WGU C951

Task 3

MACHINE LEARNING PROJECT PROPOSAL

Nicholas Ollis

Student ID #011097828

July 6th, 2023

Table of Contents

[A. Project Overview 3](#_Toc139498352)

[A.1. Organizational Need 3](#_Toc139498353)

[A.2. Context and Background 3](#_Toc139498354)

[A.3. Outside Works Review 3](#_Toc139498355)

[Sales Forecasting (Soares, 2020) 3](#_Toc139498356)

[Time Series Forecasting (Tableau, n.d.) 3](#_Toc139498357)

[Time Series Forecasting of the monthly sales with LSTM and BiLSTM (Chen, 2021) 3](#_Toc139498358)

[A.4. Solution Summary 3](#_Toc139498359)

[A.5. Machine Learning Benefits 4](#_Toc139498360)

[B. Machine Learning Project Design 4](#_Toc139498361)

[B.1. Scope 4](#_Toc139498362)

[B.2. Goals, Objectives, and Deliverables 4](#_Toc139498363)

[B.3. Standard Methodology 5](#_Toc139498364)

[B.4. Projected Timeline 6](#_Toc139498365)

[B.5. Resources and Costs 6](#_Toc139498366)

[B.6. Evaluation Criteria 7](#_Toc139498367)

[C. Machine Learning Solution Design 8](#_Toc139498368)

[C.1. Hypothesis 8](#_Toc139498369)

[C.2. Selected Algorithm 8](#_Toc139498370)

[C.2.a Algorithm Justification 8](#_Toc139498371)

[C.2.a.i. Algorithm Advantage 8](#_Toc139498372)

[C.2.a.ii. Algorithm Limitation 8](#_Toc139498373)

[C.3. Tools and Environment 8](#_Toc139498374)

[C.4. Performance Measurement 9](#_Toc139498375)

[D. Description of Data Sets 9](#_Toc139498376)

[D.1. Data Source 9](#_Toc139498377)

[D.2. Data Collection Method 9](#_Toc139498378)

[D.2.a.i. Data Collection Method Advantage 9](#_Toc139498379)

[D.2.a.ii. Data Collection Method Limitation 9](#_Toc139498380)

[D.3. Quality and Completeness of Data 9](#_Toc139498381)

[D.4. Precautions for Sensitive Data 9](#_Toc139498382)

[References 10](#_Toc139498383)

# A. Project Overview

This proposal describes a machine learning project to forecast sales. It will be a time series model, that will help predict future demand.

## A.1. Organizational Need

Currently, a sales forecast is done manually using expensive human labor and is limited to a set number of variables that current statisticians can handle. The organization wants a way to create forecasts faster and cheaper than its current human-driven process.

## A.2. Context and Background

The Ecuadorian chain store Favorita is a significant player in the retail sector, controlling multiple stores across the country. Retail stores around the world are renowned for their dynamic demand, with sales being heavily influenced by factors including, with the Ecuadorian market being additionally influenced by macroeconomic factors like oil prices. Building a reliable sales forecasting model using a time-series model could be pivotal for Favorita’s continued growth. (Alexis Cook, 2021)

## A.3. Outside Works Review

### Sales Forecasting (Soares, 2020)

This page looks at how historical data from prior sales can be used to predict expected sales. It includes code samples and explanations for different ways to solve the problem.

### Time Series Forecasting (Tableau, n.d.)

This webpage by Tableau talks about considerations that are needed to be considered for this kind of problem. Including time horizons, as well as things to check about your data to make sure you have good quality data going into the model.

### Time Series Forecasting of the monthly sales with LSTM and BiLSTM (Chen, 2021)

This article teaches concepts in predicting monthly sales numbers using an LSTM model. It includes data wrangling, transformation, and building of LSTM models for our kind of time series problem.

## A.4. Solution Summary

For this solution, I am proposing a time series forecasting model that uses advanced machine learning techniques. Considering the unique influences in the Ecuadorian market, the model will consider variables like product type, store information, promotions, and external data like oil prices and holidays. In doing this Favorita will have a model that affordably and quickly creates sales forecasts helping them better stock items and staff appropriately.

## A.5. Machine Learning Benefits

The model will improve the reliability of the model as it will not run the risk of the current human-driven system that would allow for errors to creep in at each step. Additionally, they will be able to test how promotions or other changes might affect their sales helping them plan. With this model, they will become more efficient at inventory management, operational efficiency, reduced waste, and increased profitability.

# B. Machine Learning Project Design

## B.1. Scope

* In-Scope:
  + Developing a machine learning model for sales forecasting
  + Integrating the model into an application for ease of use
  + Training and testing the model using historical data provided by the company.
* Out-of-Scope:
  + Recommendations for alterations to the physical store operations, inventory systems, and data collection systems.
  + Developing a model that auto-recommends promotions with sales forecasting.

## B.2. Goals, Objectives, and Deliverables

* Goals
  + Improve sales forecasting’s accuracy, while lowering the cost of each forecast.
  + Reduce time and human effort needed to create forecasts.
* Objectives
  + Achieve a forecasting accuracy of 90%+.
  + Reduce overstock, in addition to, out-of-stock occurrences by 50%.
* Deliverables
  + Machine Learning Model
  + Integration API
  + Application interface
  + Documentation

## B.3. Standard Methodology

We will develop will the CRISP-DM methodology.

* Business Understanding: In this initial phase we will meet with stakeholders to discuss the problem in depth. Additionally, we will review how the current system for creating the sales forecast works, and the algorithms that they use.
* Data Understanding: The dataset is to be provided by the company and contains information on sales, store details, products, promotions, and dates of transactions. Additionally, it will contain information about oil prices, holidays, and events.
* Data Preparation: We will clean and preprocess the data, making sure that missing values, outliers, and encoding categorical values are dealt with. We might also need to create new features including one that can indicate if a purchase is made on a payday or shortly after an event. We might also need to merge different datasets, like oil prices, to create a comprehensive data frame for our modeling needs.
* Modeling: We will split the data up into a training and validation set. We will start with a simple linear regression model and decision tree to establish a baseline performance. After a baseline has been set, we can experiment with more complex models such as random forest, gradient boosting, or deep learning. Time series techniques like ARIMA or LSTM will be tested for their effects on the temporal nature of our problem.
* Evaluation: We will measure the performance of the models using the metrics: Mean Absolute Error, and Root Mean Squared Error. By using both metrics we can get a good sense of not only how the model is performing but how outliers might be affecting our results. In addition to scoring, we will also check that the model isn’t overfitting to our dataset, so it might generalize well when run outside of its set data.
* Deployment: Once ready, the model will be deployed as an API for integrations with web applications and desktop applications. Depending on the model used could shift deployment details. However, we will deploy it with the recording of its predictions given the inputs so we can cross-check it with actualized sales data and make improvements when necessary.

## B.4. Projected Timeline

July 17th – The proposal is accepted, and we set a time to be onsite for a week to meet with stakeholders and learn how the current sales forecasting is being done.

July 31st – All systems needed for data review are accessible by the team and review begins.

Aug 11th – Data review complete and meeting with stakeholders to go over findings.

Sept 8th – Modeling halfway point; check-in with stakeholders about models’ performance and where the project is.

Sept 22nd – Modeling and evaluation complete. Meet with stakeholders to go over results.

Oct 6th – Project handoff and training session.

**Sprint Schedule**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** |
| 1 | July 17th | July 28th | Stakeholder meetings and process review |
| 2 | July 31st | Aug 11th | Data collection and review |
| 3 | Aug 14th | Aug 25th | Data Preparation |
| 4 | Aug 27th | Sept 8th | Modeling & Evaluation |
| 5 | Sept 11th | Sept 22nd | Modeling & Evaluation |
| 6 | Sept 25th | Oct 6th | Deployment and Documentation |

## B.5. Resources and Costs

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Resource** | **Description** | **Unit Cost** | **Units** | **Cost** |
| Lead Engineer | Technical leader | 4,000 | 12 weeks | 48,000 |
| Data Analyst | Data Review and Prep | 2,000 | 2 weeks | 4,000 |
| 2x Data Scientist | Data Prep, Modeling, and Eval | 3,500 x2 | 6 weeks | 42,000 |
| Machine Learning Engineer | Takes ML model to production | 3,500 | 4 weeks | 14,000 |
| Senior Software Engineer | Creates API, and Cloud Environment | 3,000 | 2 weeks | 6,000 |
| Project Manager | Manages communications between the two companies and the team. | 2,500 | 12 weeks | 30,000 |
| Local Hardware Cost | Prorated cost of laptops and workstations used for the project | 250 | 12 weeks | 3,000 |
| Cloud Infrastructure | Cost for setting up cloud infrastructure, and the first month of service | 1000 | 1 week | 1,000 |
|  | **Total** |  |  | 148,000 |

## B.6. Evaluation Criteria

Describe the criteria used to evaluate and measure the success of the completed project.

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Model Accuracy | The model will have an accuracy of at least 90% when tested against its test dataset. |
| Inventory Efficiency | Overstock & Out-of-stock occurrences will reduce by 50% |
| Return on Investment | The company sees gains that project their ROI for the project to be less than 3 years. |

# C. Machine Learning Solution Design

## C.1. Hypothesis

The sales of a product at a specific store can be accurately forecasted using historical sales, store information, promotional activities, and external influences like oil prices and holidays.

## C.2. Selected Algorithm

Although different types of regression algorithms will be tested. We will focus training on Long Short-Term Memory (LSTM) models, this type of recurrent neural network is good for time series forecasting.

### C.2.a Algorithm Justification

LSTM models are great for time series forecasting because they focus on what they have recently seen but can remember some things from further back. This makes them a great fit for time series forecasting. (Keith, 2022)

### C.2.a.i. Algorithm Advantage

* Temporal Dependencies
* Handling long sequences
* Flexible input lengths

### C.2.a.ii. Algorithm Limitation

* Complexity
* Prone to Overfit
* Large data requirements
* Parameter turning

## C.3. Tools and Environment

The development will take place on MacBook Pros running Apple’s M1 Max processors, with 64GB of memory. JetBrains’s tooling will be used including PyCharm Pro, Data Grip, and Webstorm. Python and Tensorflow will be used for the model creation, with pandas being used for data handling. Reports will be created inside Jupyter, and Docker will be used to isolate developer environments. The code will be hosted on GitHub.

## C.4. Performance Measurement

Data will be split in an 80/20 split into training and testing data. The model will not use any of the testing data during its training runs. The baseline of the model will be established by creating some basic regression and decision tree models. The model’s performance will be measured with Mean Absolute Error and Root Mean Squared Error on its performance when testing with the testing dataset.

# D. Description of Data Sets

## D.1. Data Source

The dataset will be provided by the client as CSV files to start. With expectations, SQL access will be offered by the time modeling has begun.

## D.2. Data Collection Method

Data will be collected manually by Favorita using the process they currently use for their current sales forecasting process.

### D.2.a.i. Data Collection Method Advantage

For Favorita, this should be easy and as they will continue to use the manual sales forecasting process until ours is deployed, it will be work that is already done each month.

### D.2.a.ii. Data Collection Method Limitation

This process will be slow and error-prone. Once we have moved the application to production, a new discussion and contract should be executed on creating an automated pipeline for the data needed for this project.

## D.3. Quality and Completeness of Data

Data will be reviewed, making sure that missing values, outliers, encoding categorical values, and binning are dealt with. Missing values will be inferred or dropped. Outliner will be removed. Encoding categorical values will be done by changing string values into integers. Some numerical values will be binned to increase the accuracy of the model.

## D.4. Precautions for Sensitive Data

As this is sales data, the potential for consumer information being in the dataset is high. For our initial project as the data is being provided from Favorita, we expect them to remove or anonymous the data so customers cannot be identified.

# References

Alexis Cook, D. i. (2021). *Store Sales - Time Series Forecasting*. Retrieved from Kaggle: https://www.kaggle.com/competitions/store-sales-time-series-forecasting/overview

Chen, D. (2021, January 14). *Time Series Forecasting of the monthly sales with LSTM and BiLSTM*. Retrieved from Denise Chen: https://denisechendd.github.io/Time-Series-Forecasting-of-the-monthly-sales-with-LSTM-and-BiLSTM/

Keith, M. (2022, January 13). *Exploring the LSTM Neural Network Model for Time Series*. Retrieved from Medium: https://towardsdatascience.com/exploring-the-lstm-neural-network-model-for-time-series-8b7685aa8cf

Soares, L. (2020, September 2). *Sales Forecasting: from Traditional Time Series to Modern Deep Learning*. Retrieved from Medium: https://towardsdatascience.com/sales-forecasting-from-time-series-to-deep-learning-5d115514bfac

Tableau. (n.d.). *Time Series Forecasting*. Retrieved from Tableau: https://www.tableau.com/learn/articles/time-series-forecasting