WGU C951

Task 3

MACHINE LEARNING PROJECT PROPOSAL

Anthony Munyan

Student ID # 001566287

June 24, 2023

Table of Contents

A. Project Overview....................................................................................................................pg 3 A.1. Organizational Need ......................................................................................................... pg 3 A.2. Context and Background.....................................................................................................pg 3 A.3. Outside Works Review ......................................................................................................pg 4 A.4. Solution Summary.............................................................................................................pg 4 A.5. Machine Learning Benefits................................................................................................pg 5 B. Machine Learning Project Design......................................................................................... pg 5 B.1. Scope ................................................................................................................................. pg 5 B.2. Goals, Objectives, and Deliverables ................................................................................. pg 6 B.3. Standard Methodology ..................................................................................................... pg 6 B.4. Projected Timeline .............................................................................................................pg 8 B.5. Resources and Costs........................................................................................................... pg 9 B.6. Evaluation Criteria ...........................................................................................................pg 10 C. Machine Learning Solution Design..................................................................................... pg10 C.1. Hypothesis.........................................................................................................................pg 10 C.2. Selected Algorithm........................................................................................................... pg 10 C.2.a Algorithm Justification....................................................................................................pg 11 C.2.a.i. Algorithm Advantage ..................................................................................................pg 11 C.2.a.ii. Algorithm Limitation .................................................................................................pg 11 C.3. Tools and Environment ....................................................................................................pg 11 C.4. Performance Measurement................................................................................................pg 11 D. Description of Data Sets .....................................................................................................pg 12 D.1. Data Source .....................................................................................................................pg 12 D.2. Data Collection Method....................................................................................................pg 12 D.2.a.i. Data Collection Method Advantage.............................................................................pg 12 D.2.a.ii. Data Collection Method Limitation ...........................................................................pg 13 D.3. Data Normalization...........................................................................................................pg 13 D.4. Data Security.....................................................................................................................pg 13 References................................................................................................................................pg 14

**Problem Statement**

Rip U. Offerton Autos, Inc. is a regional car dealership with four separate locations scattered throughout the Central Ohio area. The business specializes in selling quality, late-model used vehicles and typically carry between 300 and 400 vehicles in their company-wide inventory. The company wishes to create more support for online car shoppers on their website by helping them peruse their large, scattered inventory, which varies widely between locations. Recent trends in car-buying supports their initiative; a 2021 survey conducted by consulting firm Deloitte found that today’s car shoppers prefer a hybrid online and offline car-buying process, while only a small minority of shoppers prefer the traditional car buying process. A poll conducted by the Harris Group noted that an overwhelming majority of shoppers preferred to search for a car and determine the final price online. Thus, it seems that a robust online presence is now obligatory for car retailers, and the company has commissioned our firm to develop a ‘vehicle recommender’ for addition to their website.

The proposed Vehicle Recommender application will provide the client’s customers with a tool to enhance their online car shopping experience and increase their research productivity by:

* Asking the shopper a series of questions regarding their vehicle needs, pricepoint, and preferences.
* Recommending a vehicle from the client’s existing inventory based on the shopper’s preferences.
* Providing important attributes regarding that recommended vehicle.

The proposed application will be developed to be imbedded into the company’s website. It will access the company’s database to obtain their current inventory and its attributes. Once obtained, the information will be organized into a hash table that features the attributes of all of the vehicles in the client’s inventory. To determine the most appropriate recommendation, a decision tree will be employed to make decisions based on the association of the shopper’s preferences to the attributes of the vehicles in inventory. The shopper’s preferences will be obtained through a series of questions, which will act as decision nodes. The decision for each node will be made based on the feedback from the customer, and branches of potential vehicles that no longer apply to the customer’s preferences will be pruned immediately. The leaf (end) node will be one vehicle from inventory that matches the customer’s input. If the tree ends with more than one leaf, the algorithm will make a random selection to recommend to the user. If there is no available vehicle that matches the user’s preferences, an apology message will be generated instead.

Existing System Analysis

The client has an existing website and utilizes a cloud-based database to store information regarding their inventory. The proposed application is developed to be imbedded in their website and will have access their remote inventory database. The updates required to imbed the proposed application into their website are beyond the scope of this project.

Data

The application will retrieve current inventory data from the company’s existing inventory database, which is managed using Oracle’s MySQL. A new inventory schema will be created to featurize the attributes required for use by the application: Body Type, Make, Model, VIN, Mileage, Safety Rating, Power Source, Transmission, Drive Wheels, Occupant Capacity, Color, and Overall Condition. The application will retrieve the information from the schema in the database and parse it into a hash table to make it usable.

Project Hypothesis, Goals and Deliverables

Hypothesis:

This project will produce an application that, based on a particular customer’s needs and preferences, will recommend a vehicle from the client’s inventory to that customer. A new database schema for the client’s inventory will be developed to support the application.

Goals:

* + The application will produce a recommendation at a rate of at least 99.5%.
  + The application will make accurate decisions when associating the customer’s preferences to the client’s inventory at a minimum rate of 99%.

Deliverables:

* + A product recommendation application
  + A new inventory database schema

Project Methodology

Project development will follow the CRISP-DM methodology, which stands for Cross-Industry Standard Process for Data Mining. There will be five major phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. These phases will not be sequentially strict as it is often necessary to move between phases to develop a top-quality product. The general development of the project will follow the phases below:

* Business Understanding: The project requirements and objectives will be set according to the needs of our client. A detailed plan will be defined for each phase and for the overall product.
* Data Understanding: The quality and completeness of the client’s existing data will be examined to determine if additional field identities or reformatting is required.
* Data Preparation: Acquire additional data or reformat if necessary. Criteria for data categorization criteria will be set, as well as the importance of each attribute. During this phase, the required tables and queries will be developed.
* Modelling: The algorithm will be developed. A separate, mock dataset will be prepared for algorithm training purposes. The prepared real inventory data will be used for testing purposes. The GUI for the application will be developed. Application testing will be conducted.
* Evaluation: The results from testing will be evaluated to determine if the application meets or exceeds the project requirements. The need or desire for additional objectives will be determined.
* Deployment: When approved, the application will be released to the client’s web designer to be imbedded into the company’s website.

Funding Requirements

The estimated project costs are required resources are as follows:

* + Application Development: $48,000
  + Data Preparation: 24,000
  + Training: 5,000
  + Testing: 42,000
  + TOTAL: $119,000

B.

Problem Statement

Rip U. Offerton Autos, Inc. wants to enhance their website to make vehicle shopping easier. The business specializes in selling quality, late-model used vehicles and typically carry between 300 and 400 vehicles in their company-wide inventory, which is scattered across four separate locations. Recent trends in vehicle shopping and buying support an enhanced online presence for car dealers, and the company wants to create more support for their online shoppers on their website. By helping their customers peruse their large, scattered inventory, which varies widely between locations, the company hopes to increase sales and profitability.

The company’s current website has a listing of available vehicles for purchase, but does not filter selections based on the website visitor. The user could potentially search their entire inventory to find a particular vehicle, and it still may not meet their needs or preferences.

The proposed Vehicle Recommender application will provide the client’s customers with a tool to enhance their online car shopping experience and increase their research productivity by:

* Asking the shopper a series of questions regarding their vehicle needs, price point, and preferences.
* Recommending a vehicle from the client’s existing inventory based on the shopper’s preferences.
* Providing important attributes regarding that recommended vehicle.

The proposed application will be developed to be imbedded into the company’s website. It will access the company’s database to obtain their current inventory and its attributes. Once obtained, the information will be organized into a hash table that features the attributes of all of the vehicles in the client’s inventory. To determine the most appropriate recommendation, a decision tree will be employed to make decisions based on the association of the shopper’s preferences to the attributes of the vehicles in inventory. The shopper’s preferences will be obtained through a series of questions, which will act as decision nodes. The decision for each node will be made based on the feedback from the customer, and branches of potential vehicles that no longer apply to the customer’s preferences will be pruned immediately. The leaf (end) node will be one vehicle from inventory that matches the customer’s input. If the tree ends with more than one leaf, the algorithm will make a random selection to recommend to the user. If there is no available vehicle that matches the user’s preferences, an apology message will be generated instead.

Data Description

The client currently utilizes an on-site data server that stores data from all sales locations in once centralized location. They employ Oracle’s MySQL as their database management system, and a new schema for inventory data that excludes their customers’ personal and financial information and includes all of the necessary inventory attribute data will be developed to facilitate the proposed application’s data needs. The proposed application will access the new schema to obtain the client’s current inventory data.

The new schema will

Implementation

There will be two phases of implementation: rollout of the revised database and the release of the finalized application to the client.

Database Schema Rollout:

The appropriate personnel will be trained to make inventory changes (removing or adding vehicles, correcting errors) using the new database schema for inventory. After training, the new schema will be used concurrently with the old methodology for a period of 30 days for testing purposes. The new schema will be adopted and used exclusively for inventory purposes after the testing period.

Finalized Application Release:

After testing is complete and the application is accepted, the final version of the application will be released to the client. The client will be responsible for the implementation of the application on their website.

Evaluation

The criteria to evaluate the success of this project will be based on its goals and deliverables. At implementation, the application will be required to make recommendations, or that provide a notification that current inventory does not match their preferences, at a minimum rate of 99.5%. The decisions made by the application must be at least 99% accurate. The application will be expected to always provide the customer with the correct associative information regarding its recommendation.

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Accuracy | Makes an appropriate recommendation for 99% of customer queries |
| Associative Performance | Provides the customer with a recommendation, or that current inventory does not match their preferences, for 99.5% of customer queries |
| Data Retrieval Accuracy | The data retrieved regarding a recommendation is ALWAYS the correct associated data |

Testing will consist of three methods: unit testing, usability testing, and acceptance testing. Throughout project development, unit testing will be employed to ensure proper function of all parts of the application. Usability testing will be conducted after each development unit is finished. Acceptance testing will be performed after development is complete but before the final release of the application.

Resources and Costs

The projected costs are required resources are as follows:

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| MySQL  (Data Analytics) | Existing database management system | $0 |
| Recommendation Application  (Developers) | Development of recommendation algorithm and data-wrangling software and associated GUI | 48,000 |
| Data Preparation  (Data Analytics) | Extraction, processing, cleaning, and preparation of data, including new schema for inventory; create mock data set for training algorithm | 24,000 |
| Inventory Management Training  (Management) | Training appropriate personnel on making inventory changes using new schema | 5,000 |
| Testing  (Developers) | Training and testing the application using mock data and real data | 42,000 |
|  | **Total** | ($119,000) |

Project Timeline

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Event** | **Start** | **End** | **Dependencies** | **Tasks** |
| Evaluation and Preparation Phase | January 3, 2024 | January 12, 2024 | None | Evaluate data, cleanse, prepare, and generate data for training and testing |
| Inventory Schema Development and Data Preparation | January 15, 2024 | February 2, 2024 | Completion of data evaluation phase | Prepare and generate data for training and testing, develop new schema for product inventory |
| Personnel Training | February 5, 2024 | February 9. 2024 | Completion of data prep and schema development phase | Train appropriate personnel to use new inventory schema |
| Database Testing | February 12, 2024 | March 15, 2024 | Completion of database training phase | Test the new database schema |
| Machine Learning Algorithm Development | January 20, 2024 | March 8, 2024 | Completion of database training phase | Develop the machine learning algorithm and GUI |
| ML Training | February 12, 2024 | March 8, 2024 | ML algorithm must be developed | Train the machine learning algorithm |
| Application Testing | March 8, 2024 | March 22, 2024 | Machine learning training phase must be complete | Testing the application |
| Release | March 25, 2024 | March 27, 2024 | Application must be accepted by client | Release finished product to client for deployment |

**A.4. Solution Summary**

The application will be for recommendation purposes only and will not determine an individual’s or businesses’ actual worthiness for a mortgage. It will not prequalify a customer for, qualify a customer for, or grant a customer a mortgage.

D4 Hypothesis Verification

The hypothesis that an application could be developed to provide loan officers the most appropriate mortgage for a potential customer, based on that potential customer’s information and how it matches the attributes of available mortgage packages, has proven to be correct. Test cases have yielded an appropriate recommendation for 95% of the testing data set. The decision tree prescriptive method has proven to be an excellent algorithm for providing a single recommendation in this instance, which is the primary objective of the application. Future iterations of the application should provide at least one alternative recommendation to better serve potential customers, which could be achieved by developing multiple decision trees or a weighted scoring system.

Accuracy Analysis

Accuracy was measured using cross validation. For each test case, evaluators were given a specific set of answers to enter into the application, as well as the appropriate recommendation the application should make based on those specific answers. If the application provided the appropriate recommendation, the test passed. If any recommendation other than the appropriate answer was provided, the application failed the test. The pass rate was over the required threshold of 95%.

.

D8 Application Files

D9 User Guide

For evaluation of the application,

1. The application was developed using Python 3.7 and JetBrains Pycharm. It is recommended that a similar environment be used for evaluating the application.
2. Unzip the provided files into a new folder.
3. Start the application
4. Make selections for each question using the check boxes or radio buttons provided in the GUI.
5. Left click on the ‘ENTER’ button and observe the result
6. To exit the application, left click on the ‘DONE’ button or close the GUI window.

**A.5. Machine Learning Benefits**

The proposed chatbot’s utilization of machine learning to analyze customer inquiries will reduce the amount of human intervention needed to address basic inquiries. It will forward more advanced inquiries to the appropriate customer service personnel that specializes in that product or category, which in turn will have more time to address their concerns. By categorizing complaints, the average response time will be significantly reduced and customer satisfaction with the resolution process will significantly increase. Faster response times and better resolutions will provide a positive experience and increase value to our customers.

**B. Machine Learning Project Design**

**B.1. Scope**

* An online chatbot that handles and analyzes online customer inquiries and complaints will be developed.
* A dashboard to analyze the performance of the chatbot will be created.
* Our web server will be upgraded to handle increased traffic and the increase in resource demands created by the chatbot.
* The proposed improvements do not include a provision for telephone calls to customer service. They will continue to be handled in the current fashion.

**B.4. Projected Timeline**

|  |  |  |  |
| --- | --- | --- | --- |
| **Event** | **Start** | **End** | **Tasks** |
| Evaluation and Preparation Phase | August 1, 2023 | September 30, 2023 | Evaluate current customer service practices and personnel, evaluate data, Cleanse and prepare data for training and testing |
| Development | August 1, 2023 | September 30, 2023 | Develop the code for the chatbot and performance dashboard |
| ML Training | October 1,2023 | October 31, 2023 | Training the chatbot |
| Testing | November 1, 2023 | November 30, 2023 | Testing the chatbot and dashboard |
| Personnel Training | November 15, 2023 | December 1, 2023 | Re-assign customer service personnel to appropriate teams if necessary, train personnel on the upgraded system |
| Deployment | December 1, 2023 | December 15, 2023 | Deployment |

**B.5. Resources and Costs**

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Web Server  (IT) | Upgrade due to resource cost of chatbot and associated software | $20,000 |
| RabbitMQ  (Developers) | Open source, but will be tailored by development team to fit our needs | 8,000 |
| Kubernetes API/ Kubernetes  (Developers) | Open source, but will need to be tailored by the development to fit our needs | 48,000 |
| PostgresSQL  (Data Analytics) | Existing database software, software may need tailored for additional duties | 1,000 |
| Machine Learning Engine  (Developers) | Development of the machine learning algorithm and associated software | 48,000 |
| Data Preparation  (Data analytics) | Extraction, processing, cleaning, and preparation of data set | 42,000 |
| Chatbot and Dashboard Training/ Testing  (Development) | Training the chatbot using data set, testing chatbot and dashboard | 40,000 |
| Develop Inquiries and Complaint Categories  (Management) | Develop the categories of complaints and inquiries that will be used by customer service | 20,000 |
| Personnel Training  (Management) | Train personnel on changes | 30,000 |
|  | **Total** | ($257,000) |

**C. Machine Learning Solution Design**

**C.1. Hypothesis**

Within three months of deployment, the proposed chatbot will adequately handle 90% of minor inquiries and complaints without human intervention. Major concerns or complaints will be forwarded to the proper customer service team 90% of the time.

**C.2. Selected Algorithm**

Random Forest, a supervised machine learning algorithm, will be used to power the chatbot. Supervised machine learning uses pre-categorized data to train algorithms for classification tasks. Random Forest combines the predictions of multiple individual models to make a final decision. Each individual model consists of its own decision tree that makes a decision based on input data.

**C.2.a Algorithm Justification**

Random Forest algorithms are effective in processing natural language inputs to make accurate predictions about their intent and meaning. The algorithm is especially useful for classifying user inputs into different categories, like requests to complete a simple task, for assistance, or for information. These strengths align well with the goals of this proposed project.

**C.2.a.i. Algorithm Advantage**

Random Forest algorithms are known to be stout algorithms that can handle noisy data and missing values. They are particularly stable and accurate.

**C.2.a.ii. Algorithm Limitation**

Random Forest algorithms are highly complex and require a longer training period, which results in higher development costs compared to other machine learning algorithms.

**C.3. Tools and Environment**

The chatbot messages will be handled with RabbitMQ, which is message broker software that will be loaded onto our web server. To handle multiple users, the chatbot platform will be equipped that the ability to rapidly scale. This entails an architecture to create separate containers for each instance of the chatbot that is in service. Kubernetes API and Kubernetes open-source software will be used to containerize each session of the chatbot, and balance load demands between them. The Random Forest algorithm that powers the machine learning abilities of the chatbot, the associated dialog, and analytical software will be written in Python and stored on our web server. PostgreSQL database management software will process and store chatlogs on our current database server.

**C.4. Performance Measurement**

Quality will be measured according to the accuracy of the chatbot’s predictions on the categories of minor user inquiries and solutions to minor user complaints. The chatbot’s categorization and selected user team to deal with major inquires and complaints will be used as well. Quality will be assessed weekly by the data analytics department, who will randomly sample logs from that week to make an assessment. Performance will be measured by tracking the time it takes the chatbot to resolve problems or recommend a customer service team to solve a problem; and by tracking the percentage of minor inquiries and complaints it handles compared to the percentage that gets forwarded to live customer service. The performance dashboard will track and display real-time performance results.

**D. Description of Data Sets**

**D.1. Data Source**

Data for this project will be sourced from logs and transcripts from our own company’s past customer service events, which consist of transcribed telephone calls and emails.

**D.2. Data Collection Method**

This project will require at least 10,000 complete, analyzed, and cleaned datapoints from past customer service events. Information pertaining to past customer service events is located on the company data server, which currently contains records of 50,000 past events. The age of the data ranges from new (collected last week) to ten years old. Our data analytics team will retrieve and analyze the data for completeness before cleansing and categorization. To avoid confusion and corruption of our existing data, the raw data used for this project will be copied and moved to a different storage location before processing.

Our customer service guidelines will be used to determine the completeness of past inquiries and complaints. Inquiries that have been answered accordingly will be considered complete. Likewise, complaints that have a resolution will be considered complete. Incomplete, illegible, or unintelligible inquiries and complaints will be cleansed from the data set.

Senior managers from customer service will be tasked with creating categorization criteria for the customer service inquiries and complaints. Once such criteria are established, data analytics will pore through the remaining data points and categorize them accordingly.

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

Using our own customer service data is advantageous for both our data analytics team and our machine learning development team. Data analytics won’t be burdened with weeding out data points that have no relevance to our company or this project, while the development team will be provided with a wealth of data that relates to our own products and business practices and will be ideal for training and testing the machine learning algorithm.

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

Our stored customer service data is raw and will require a lot of processing to prepare it for use for this project. Another challenge is the volume of raw data and the volume of prepared data that is required for this project.

**D.3. Quality and Completeness of Data**

Once categorized, the data will be prepared for training and testing. This will involve transforming the data into a format and file structure that will be usable by the machine learning algorithm. Transformation will include, but is not limited to, the following steps: changing the text case to all lowercase, removing insignificant words (‘and’, ‘the’, ‘a’,), removing meaningless punctuation marks, and removing excess white space, and converting the files to the proper file format.

**D.4. Precautions for Sensitive Data**

Users will be encouraged to avoid using personal information in the chatbot. Order numbers, transaction numbers, product numbers, or product descriptions will be used in lieu of personally identifiable information. A provision for identifying and scrubbing personally identifiable information will be developed.

Testing consisted of three methods: unit testing, usability testing, and acceptance testing. Throughout project development, unit testing was employed to ensure proper function of all parts of the application. Usability testing was mainly employed after each development unit of the application was finished. Acceptance testing was performed after primary development of the entire application was complete.

Unit Testing

Unit testing was used to test all units of the application, including data wrangling, the functionality and accuracy of the decision tree algorithm, and the functionality of the GUI.

Usability Testing

Testers were asked to perform tasks using the application and provide feedback based on overall performance, usability, and any bugs they may have encountered. Examples of test cases included attempting to not answer questions or attempting to provide multiple answers to the same question.

Acceptance Testing

Stakeholders reviewed the features and functionality of the application against the project requirements. Once all project requirements were met, the application was tested for bugs, accuracy, and performance

**References**

Morgeson III, F. V., Hult, G. T., Mithas, S., Keiningham, T., & Fornell, C. (2020). Turning Complaining Customers into Loyal Customers: Moderators of the Complaint Handling–Customer Loyalty Relationship. *Journal of Marketing*, 79-99.

Ramya, K. R., Chandrika, D., Nikhitha, K., Niteesha, T., & Pravallika, P. (2021). Opinion Mining Using Social Media. *Turkish Journal of Physiotherapy Rehabilitation*, 20-30.

Statista. (2018). *Expected Response Time for Social Media Questions or Complaints in US and Global 2018*. Retrieved from Statista: https://www.statista.com/statistics/808477/expected-response-time-for-social-media-questions-or-complaints/

Wang, Y. (2022). Using Machine Learning and Natural Language Processing to Analyze Library Chat Reference Transcripts. *Information Technology & Libraries*, 1-10.

Zota, R. D., Cimpeanu, I. A., & Dragomir, D. A. (2023). Use and Design of Chatbots for the Circular Economy. *Sensors*, 4990-5010.