Data Analytics Capstone Topic Approval Form

The capstone challenges students to integrate skills and knowledge from several program domains into one project. The guidelines for this capstone course require you to demonstrate the application of academic and professional abilities developed as an undergraduate student in the BSDA program. It is highly recommended that the topic of your capstone be about resolving a current or perceived business problem. Your research topic should exemplify scholarship and research at the highest level and should be significant enough that it would help potential employers identify your abilities. It is also recommended that you use publicly available datasets for transparency and external validity.

This document is designed to help you clearly state the research question you will be exploring in your capstone project, the scope of your project, and your timeline in order to ensure that all of these align with your degree emphasis. Without clearly defining each of these areas, you will not have a complete and realistic overview of your project, and it cannot be accurately assessed whether your project will be acceptable for this capstone course.

If your project is one you have already completed at work or elsewhere, this document should be easy to complete. Many students do use a project they have already completed in the past. In that case, you will write the proposal as if the project has not been completed yet, and when you report on your project, you will use your complete after-implementation report. If you have not yet completed your project, this document can help ensure the scope is within the acceptable range for this capstone. An instructor must approve this form before you submit this task for evaluation. The task will not be evaluated without an instructor's signature. The instructor may ask for additional information before approving this form.

Before submitting this form for approval, please remove all italicized directions in the form.

Please only submit a Topic Approval Form that has been signed by a course instructor for evaluation.

Capstone Topic Approval Form

The purpose of this document is to help you clearly explain your capstone topic, project scope, and timeline. Identify each of the following areas so you will have a complete and realistic overview of your project. Your course instructor cannot approve your project topic without this information.

Note: You must fill out and submit this form. Space within each section will expand as needed.

Any costs associated with the development of the data analytics solution will be the responsibility of the student.

Student Name: Andrew Perez

Student ID: 010903754

Capstone Project Name: Optimized Used Car Pricing with Machine Learning

Project Topic: Creating a random forest model that can accurately predict the optimal sales price for a used car using a variety of predictors.

Research Question: Can a random forest model be fitted to predict the optimal sales price of a used car with a Root Mean Squared Error (RMSE) value that is less than 5% of the dataset's average price?

> **Hypothesis:** The random forest model can predict pricing while maintaining an RMSE value of less than 5% of the dataset's average price.

Context: Precision in pricing is pivotal for a dealership's success. Introducing a machine learning model that not only fine-tunes pricing but also minimizes errors can revolutionize operations. The suitability of a random forest model for this task is evident. It adeptly manages the intricate relationships and nuanced interactions in the automotive sector's diverse set of specifications.

Data: To adequately test this hypothesis, I need to collect a used car dataset that has a variety of vehicle specifications and is sizeable enough to draw a significant conclusion. The dataset I plan on using can be found here:
https://www.kaggle.com/datasets/alihassankp/dubizzle-used-car-sale-data

> *If an existing dataset will be used, describe the dataset:* The dataset has 9170 entries and 20 field types. Most of the field types are 100% complete and of the fields that are not, only 1 has 10%+ of its data missing. The data was scrapped from a United Arab Emirates marketplace called Dubizzle and it was last scrapped on 12/05/2022.
>
> *Explain who owns the data and why you are allowed to use the data for your capstone project:* The data is publicly available on Kaggle and is free to use.

Data Gathering: *Describe the data-gathering methodology you will use to collect data.* I will download tabular data from Kaggle.

Data Analytics Tools and Techniques: I will perform a regression analysis on this dataset using a random forest model and potentially analyze feature importance if the model can accurately predict a used car's optimal price. I will be using Python and its various libraries, including sci-kit-learn and pandas, to perform these analyses as well as any cleaning and/or manipulation that is required.

Justification of Tools/Techniques: Choosing a random forest model makes sense for our research question for a few key reasons. First, it's good at handling complicated relationships between factors, which is crucial given the many features affecting used car prices. Plus, it's flexible enough to deal with both numbers and categories, and our dataset has a mix of both. Lastly, the model comes with a built-in feature ranking, making it easy to see which features matter the most if our model works well and rejects the null hypothesis.

Application Type, if applicable (select one):

- ☐ mobile
- ☐ web
- ☒ stand-alone

Programming/Development Language(s), if applicable: Python

Operating System(s)/Platform(s), if applicable: N/A

Database Management System, if applicable: N/A

Project Outcomes: *List the key anticipated project outcomes and deliverables in fewer than 500 words*. The project's focus is on utilizing data analysis and machine learning to improve pricing strategies for used cars, with the following outcomes/deliverables:

The first goal is to create an accurate pricing model through regression analysis. This model, utilizing factors like kilometer count, horsepower, and year, aims to accurately predict optimal sales prices for used cars.

Next, to ensure the reliability of the regression model, the project includes an outcome focused on model validation and accuracy assessment. This involves presenting metrics and visualizations demonstrating the model's performance, including measures like Root Mean Squared Error (RMSE) and visual comparisons of predicted versus actual prices.

Lastly, the project aims to optimize pricing strategies for dealerships by offering actionable insights with the assistance of feature importance ranking. This analysis identifies key factors influencing pricing decisions, empowering dealerships to refine their strategies.

Projected Project End Date: 12/31/2023

Sources:

1) Dubizzle Used Car Sales Data (https://www.kaggle.com/datasets/alihassankp/dubizzle-used-car-sale-data).

2) Random Forest: A Complete Guide for Machine Learning (https://builtin.com/data-science/random-forest-algorithm#feature).

3) ML Driven Dynamic Pricing @ CARS24 – Part 1 (https://medium.com/cars24-data-science-blog/how-cars24-uses-machine-learning-for-dynamic-pricing-of-used-cars-part-1-51fee52860d1).

Human Subjects or Proprietary Information

Does your project involve the potential use of human subjects? (Y/N): N
Does your project involve the potential use of proprietary company information? (Y/N): N

STUDENT SIGNATURE

Andrew Perez

By signing and submitting this form, you acknowledge that any cost associated with the development and execution of your data analytics solution will be your (the student's) responsibility.

TO BE COMPLETED BY AN INSTRUCTOR

The capstone topic is approved by an instructor.

COURSE INSTRUCTOR SIGNATURE:

Jim Ashe, Ph.D. Mathematics

COURSE INSTRUCTOR APPROVAL DATE:

12/8/2023

Project Compliance with IRB (Y/N): Y