November 30, 2023

C964: Computer Science Capstone

Part A: Letter of Transmittal 1

[Letter of Transmittal Requirements 2](#_Toc1738085866)

[Letter Template 2](#_Toc1133266111)

[Part B: Project Proposal Plan 3](#_Toc1370766476)

[Project Summary 4](#_Toc904507251)

[Data Summary 4](#_Toc1736393957)

[Implementation 4](#_Toc1241988654)

[Timeline 4](#_Toc1357178365)

[Evaluation Plan 5](#_Toc623361460)

[Resources and Costs 5](#_Toc1538507987)

[Part C: Application 5](#_Toc1471073175)

[Part D: Post-implementation Report 6](#_Toc651895932)

[Solution Summary 7](#_Toc1134136520)

[Data Summary 7](#_Toc182221765)

[Machine Learning 7](#_Toc1505466430)

[Validation 7](#_Toc391434166)

[Visualizations 7](#_Toc201059345)

[User Guide 7](#_Toc1365484010)

[Reference Page 8](#_Toc1702353417)

# Part A: Letter of Transmittal

## Letter of Transmittal

11/29/2023

John Jameson, CTO

Auto Auctions Online, Inc.

100 Motor Way

Chicago, IL 60106

Dear Mr. Jameson,

Last quarter, our company leadership laid out several high-level objectives that we, as a company, are aiming to achieve in the next year. My team was tasked with enhancing the customer user experience on our website. I know you will agree that as the leading automotive auction website in the United States, it is imperative that we continue to exceed customer expectations by employing innovative technology to provide an excellent customer experience which in turn helps us retain our loyal customers as well as attract new customers. We have two types of customers to satisfy – sellers and buyers. Sellers need to be able to quickly list their automobiles for auctions, and buyers need to be able to quickly find the type of vehicle they are seeking to buy. As you will see, the correct categorization of vehicle type is instrumental to both of these use cases.

The solution I am proposing today is a convolutional neural network – in short it is a technology that will identify and categorize an image that is uploaded by a customer. This categorization will automatically be applied to customer listings so that sellers won’t need to select the category manually, saving them time and improving listing accuracy. Buyers will be able to filter by vehicle type that was provided through this automated process, which improves the ease of use on the website, and also ensures that the customer will see the type of automobile listings that they are seeking.

With our proposal in mind, we seek to increase revenue by driving more sales. Buyers can spend less time finding the vehicle that they desire to buy with improved listing accuracy and dynamic filtering provided by accurate seller listings and automated categorization. We will improve customer retention rates and customer satisfaction scores by enhancing the user experience on the website. Because listing accuracy will be improved through the use of the convolutional neural network, we will reduce the overall number of customer service requests, and reduce the number of support cases received from customers.

Please find my proposed timeline for this project below. I am estimating that my team can have this solution implemented within four weeks. A project manager and I will lead a scrum team of three developers over two sprints which we will prioritize for next month, with your approval. I will coordinate with our customer support leaders to ensure that knowledge articles are updated and customer support representatives receive appropriate training on the changes. The only additional thing we will need for this project is a Google Colab Pro+ monthly subscription.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sprint 1** | | **Sprint 2** | |
|  | **Week 1** | **Week 2** | **Week 3** | **Week 4** |
| Build CNN |  |  |  |  |
| Train CNN |  |  |  |  |
| Test CNN  CSR Training  KB Updates |  |  |  |  |
| Deploy CNN |  |  |  |  |

John, I have had the pleasure of reporting to you during my tenure at the company, and I know you will carefully consider my proposal for its own merits, but I would also ask you to consider my record of success in implementing new technology that has improved the customer experience as well as increased the company’s revenue. I have brought my passion for neural networks from my previous employment, and while this proposal could be considered a quick win, this is just the beginning of how we can utilize this technology to benefit the company – there are countless more options for automating the listing process that could be considered in future iterations. With your support behind this project, I am confident that leadership will agree to move forward with this solution. I appreciate your consideration and eagerly await your response.

Sincerely,

Andrew Stowe

Senior Software Engineer

# Part B: Project Proposal Plan

The project proposal should target your client’s middle management. This audience may be IT professionals but have limited computer science expertise. Use appropriate industry jargon and sufficient technical details to describe the proposed project and its application. Remember, you’re establishing the technical context for your project and how it will be implemented for the client. **Write everything in the future tense.**

## Project Summary

* Describe the problem.
  + Sellers may manually enter a category for the vehicle type, but it is not required, and also introduces the opportunity for user error.
  + Automotive listings are not able to be accurately filtered by vehicle type on the website.
* Summarize the client and their needs as related to the problem.
  + Auto Auctions Online, Inc. is seeking to improve the customer experience on the website.
* Provide descriptions of all deliverables. For example, the finished application and a user guide.
  + Convolution Neural Network build
  + Convolutional Neural Network training
  + Staging on testing website
  + Testing on the testing website
  + Deployment to production website
  + Change management support
  + Updates to knowledge base articles
* Provide a summary justifying how the application will benefit the client.
  + Auto Auctions Online, Inc. will increase revenue by driving more sales. Buyers will spend less time searching for the vehicle that they wish to buy. Customer retention rates and customer satisfaction scores will be improved by enhancing the user experience on the website. The overall number of customer service requests will be reduced, as will the number of support cases received from customers for mislabeled listings.

## Data Summary

* Provide the source of the raw data, how the data will be collected, or how it will be simulated.

Training data will be acquired from kaggle.com at the following link: <https://www.kaggle.com/datasets/ryanholbrook/car-or-truck>

* Describe how data will be processed and managed throughout the application development life cycle: design, development, maintenance, or others.

The training data will be read into the program after which it will be converted to grayscale images. Every image in the dataset is 128 by 128 pixels and will be left at this size for training purposes. After all the images are in the system they will be split into a training set(data used to train) and a validation set(data used to test). The training data will then go through a neural network which will classify each image and the accuracy of its predictions will be recorded. The model resulting from that run-through with the training data will then be used to predict what each image in the validation set is, and the resulting accuracy will be recorded. This training and validation process will occur between ten and twenty times to produce a reasonably accurate model.

After the model is produced there will be no more required processing of data. One thing to consider however is that the model is being specifically trained on square images. This means that images that do not have a square aspect ratio would be potentially distorted and misclassified if they are fed into the model. There is an easy workaround to this in that we can require users to crop all of the images they upload into square proportions.

* Justify why the data meets the needs of the project. If relevant, describe how data anomalies, e.g., outliers, incomplete data, etc., will be handled.

The data meets the needs of the project because it includes many images of cars and trucks that are already categorized as cars or trucks. One thing to consider is that all customer images will need to be cropped to be square in proportion as the model is being trained on square images.

* Address any ethical or legal concerns regarding the data. If there are no concerns, explain why.

There are no ethical or legal concerns at this time. Vehicle type is publicly accessible information with no implications to privacy or ethical concerns.

## Implementation

* Describe an industry-standard methodology to be used.

The project will follow the scrum methodology and will be completed over two sprints with daily scrum meetings.

* An outline of the project’s implementation plan. This outline can focus on the project’s development as a whole, or it may focus on only the implementation of the machine learning solution.

The first step will be to acquire the dataset from Kaggle.com. After this, we will process the data so it can be used to build a model using Jupyter Notebook and Visual Studio Code. Next, we will build the model using TensorFlow. Once the model is built, we will train the model on the processed dataset. When the model is trained it can then be assessed for accuracy and saved. The saved model can then be deployed to a Google Colab notebook in order to make predictions.

## Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected start date | Anticipated end date |
| Trained Model | 14 days | 11/1/2023 | 11/15/2023 |
| Google Colab Notebook | 15 days | 11/15/2023 | 11/30/2023 |

## Evaluation Plan

* Describe the verification method(s) to be used at each stage of development.

The verification method that will be used is integrated into the development of the convolutional neural network in the form of the accuracy metrics from the validation data set. This information will tell us how accurate the model is at predicting whether images in the validation set are cars or trucks.

* Describe the validation method to be used upon completion of the project.

Customer Support Managers will monitor customer support requests to ensure that functionality issues are not being reported in relation to the implementation. If images are being misclassified in large amounts this will create a large amount of customer support tickets which will allow us to reassess the model.

## Resources and Costs

* Itemize hardware and software costs.

Hardware Costs:

Computer for development tasks: Free (we already own this)

Software Costs:

Visual Studio Code: Free

TensorFlow: Free

Jupyter Notebook: Free

Google Colab Pro+: $49.99 per month

* Itemize estimated labor time and costs.

Project Manager:

19 working days \* 1 hour per day = 19 hours \* $40/hour = $760.00

Lead Software Engineer:

19 working days \* 8 hours per day = 152 hours \* $50/hour = $7,600.00

Developer 1:

19 working days \* 2 hours per day = 38 hours \* $35/hour = $1,330.00

Developer 2:

19 working days \* 1 hour per day = 19 hours \* $35/hour = $665.00

Developer 3:

19 working days \* 1hour per day = 19 hours \* $35/hour = $665.00

Total estimated hours: 247

Total estimated labor cost: $11,020.00

* Itemize estimated environment costs of the application, e.g., deployment, hosting, maintenance, etc.

Google Colab Pro+: $49.99 per month

# Part C: Application

<https://colab.research.google.com/drive/1PvRCbGjUpIWgGGSzwysdZUvk29ZzUCd9?usp=sharing>

**Part D: Post-implementation Report**

## Solution Summary

* Summarize the problem and solution.

The problem in this scenario was a need to distinguish between cars and trucks for an automobile auction website. The solution was a convolutional neural network that can predict whether something is a car or a truck based on an input image.

* Describe how the application provides a solution to the problem from parts A and B.

The application allows users to predict whether an image is a car or a truck with a simple command typed into and run on a Jupyter Notebook on Google Colab. It can predict based on a direct image URL or an image file. This will allow for the easy identification of user-submitted images.

## Data Summary

* Provide the source of the raw data, how the data was collected, or how it was simulated.

The data for the model was acquired from <https://www.kaggle.com/datasets/ryanholbrook/car-or-truck>.

* Describe how data was processed and managed throughout the application development life cycle: design, development, maintenance, or others.

The data had a fairly minimal amount of manipulation done before it was used to train the model. It was read into an array with category labels at the size it came in, which was 128x128. It was converted to a grayscale, as color is not a relevant factor in distinguishing cars from trucks. Next, the data was put in a random order to reduce bias during training. The image array data was then normalized by dividing it by 255 to reduce the computation time when it was being run through the model. After this, it was finally run through a convolutional neural network with three convolutional layers for ten epochs to generate the final model.

## 

## Machine Learning

For each employed method (at least one is required) provide the following:

* Identify the method and what it does (the “what”).

The machine learning method used for this project was a convolution neural network(CNN). A CNN is a type of deep learning architecture that learns directly from data. Each layer in a CNN learns to detect different sets of features in an image. Each layer then passes its filtered image onto the next layer which then applies a new set of filters. After going through all of the convolutional layers the images go to a dense layer which will classify the images based on the probability of the possible classes, in this case, a truck or a car.

* Describe how the method was developed (the “how”).

The method was developed using a Jupyter Notebook, TensorFlow, and Python in Visual Studio Code. The parameters of the convolutional layers were iterated over time in order to balance the speed at which the model was trained and the final accuracy of the model. The goal was to train the model in less than 10 minutes and to have a validation accuracy greater than 80%.

* Justify the selection and development of the method (the “why”).

A CNN was selected for this use case because they are ideally suited to the task of image recognition. They have the capability to reduce the number of parameters of a dataset without diminishing the quality of the model. The language Python was chosen because it has many libraries which makes doing machine learning tasks much simpler. Jupyter Notebook was used because it allows the piecemeal execution of code which enables swift iteration and tuning when dealing with a neural network. TensorFlow was selected for use because it is free and ideally suited to working with machine learning problems involving convolutional neural networks.

## Validation

For each employed method described in the section above provide the following:

* An appropriate validation method.

The convolutional neural network developed with TensorFlow is self-validating. When training the model 20% of the available data was isolated to validate the accuracy of the resulting model of each training epoch.

* Results of the validation method or a plan to obtain those results.

The results of the validation method can be viewed in the submitted Google Colab by navigating to the cell that is fourth from the bottom labeled TensorBoard. This cell contains a graph called epoch-accuracy which shows how the accuracy of the model increased over the course of training. One of the lines in this graph is validation accuracy which can be identified using the key on the left side of the output cell.

## Visualizations

All three visualizations can be generated at the Google Colab link by following the instructions in the user guide. They occupy the fifth from the bottom, fourth from the bottom, and third from the bottom code cells.

## User Guide

1. Open the following link: [Google Colab](https://colab.research.google.com/drive/1PvRCbGjUpIWgGGSzwysdZUvk29ZzUCd9?usp=sharing)
2. Click the folder icon in the left margin of the window Google Colab opened in
3. Upload the zip files (carlogs.zip, carstrucks.model.zip, vehicle2.zip) by dragging them from File Explorer to the margin that opened when you clicked the folder icon (you should see something pop up that says “Drop files to upload them to session storage.”)
4. Wait for the files to finish uploading (this might take a few minutes, you can tell by the circles that appear)
5. Select the Run all option from under Runtime on the menu bar and click Run anyway in the warning box that pops up(this might also take a few minutes to complete).
6. Navigate to the bottom of the page where in the bottom five code cells you will see the three visualizations in addition to two cells making example predictions.
7. These example prediction cells can be edited and run again to make new predictions by following the instructions located above them.