CarEnvision: A Data-Driven Machine Learning Framework for Automated Car Value Prediction

TianGe (Terence) Chen[1], Angel Chang[1], Evan Gunnell[2], Yu Sun[2]

[1]Rancho Cucamonga High School, Rancho Cucamonga, CA, 91701
[2]California State Polytechnic University, Pomona, CA, 91768

Abstract

When people want to buy or sell a personal car, they struggle to know when the timing is best in order to buy their favorite vehicle for the best price or sell for the most profit. We have come up with a program that can predict each car's future values based on experts' opinions and reviews. Our program extracts reviews which undergo sentiment analysis to become our data in the form of positive and negative sentiment. The data is then collected and used to train the Machine Learning model, which will in turn predict the car's retail price.

Keywords

Machine Learning, Polynomial Regression, Artificial Neural Network.

1. Introduction

The struggle of buying a new car can also bring the conflict of knowing when to sell it for maximum profit. Thousands of car owners ranging from multiple age groups have this problem. According to I. Wagner, in 2019, around 91.3 million motor vehicles were sold globally [1].Nonetheless, the key component in finding the perfect time is the popularity and the voice of previous car owners over a series of years. This thought occupied us for the majority of our project. How are we going to extract millions of data values to help car owners around the globe? Through the use of top-of-the-line technology, we were able to bring the benefits of CarEnvision for everyone's usage. Without CarEnvision, thousands of car owners are hit with the question of: is this the right time? CarEnvision allows car owners to answer this simple question. However, the car industry changes every day and on, making this assumption quite difficult. So, the solution? CarEnvision. It is easy to use through the use of machine learning and automation [9, 10, 11]. With a short simple survey on the car, a single press of a button can tell you if it is the right time. It gathers thousands of information from the vast interweb to determine the price of the car next year. With this information, it can then help you decide if it's the right time to sell the car. The difference between profit and no profit can determine the future of the car industry.

Throughout the car market industry, many car predicting software have been presented to the public to help ease the struggle of selling your own car. These software are engineered to use the pricing of car dealerships without regard to the profit, policies, or managers that determine the price points. Some dealerships may have had different amounts of customer interaction and therefore have fewer or more sales. This can greatly impact the companies' choice of the price of cars. With changing price tags depending on the dealership, the value of the car may not be stable

David C. Wyld et al. (Eds): COMIT, CRBL, BIOM, WiMNeT, SIP, AISO - 2021
pp. 17-25, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.111702

and not be reliable. In addition, constant bargaining for cheaper price points may also affect the value of the car. According to Edmunds, “the more car deals the car salesperson makes, the more money that salesperson takes in.” The pricing of the car uses different factors such as the manufacturer’s price. With hopes of gaining maximum profit, dealerships may input policies or change the price tag. A second practical problem is that the car value predictor may not use the opinions of the buyers. When selling a car, knowing that the population has a positive view on this model can increase the value of the car. Nonetheless, the most important aspect of selling a car is the opinion of the buyer. As a result, our program, CarEnvision, allows for users to easily extract information that is not artificially produced, but straight from the opinions of hundreds of car owners.

Though CarEnvision is unique, it is not the only predictor in the field of artificial intelligence. For example, Market Insider is a website that collects data from the stock market, and creates a two-dimensional graph depicting the growth and decay of car values over time. It also provides live changes of every company’s stock values whether they are dropping or climbing. But our program, CarEnvision, assembles big data together from a public car website, analyzes the sentiments to see how positive or negative the opinions are, and uses Machine Learning algorithms to predict a future car value. This prediction is not a feature of Market Insider as they only provide the historical activities of the car company’s values. Our goal is for the program to predict an accurate retail price for the car in the upcoming year using enough collected information. That is why we used jdpower.com as our main source to gather the chosen car’s past five years of public opinions and reviews. This data is able to then train our Machine Learning model to be as accurate as possible. With more and more data provided, the predicted retail price will fluctuate accordingly which is why our program provides trusted information.

In order to keep our data up-to-date and reasonable, we relied on jdpower.com. This website includes universal information to almost every old and new car on the market. Whenever new cars are hinting to hit the market, new data will begin to appear on the website for users to see. Using the first part of our algorithm code, the chosen car’s opinions are downloaded directly off of the website itself. For example, each opinion is evaluated, producing a level of positivity and negativity. The sentiments are used to train the Machine Learning Model. For every new prediction, the program pulls data live off of the website. These are examples of different car models we personally chose which input the sentiments into our model. We used four different car models with four different sets of sentiments which make our prediction as accurate as possible. Furthermore, the output of our model would be the car’s value which is the quotient of the current price (used) divided by the release price.

The added paper’s structure is organized in different sections that include the process and details about CarEnvision. Section 2 is about the challenges we faced during our development and procedure. Section 3 provides the solution we came up with in order to resolve the conflicts mentioned in section 2. Section 4 includes in-depth explanation on how our program was created and the reason behind the goal, along with a presentation that connects to our work in section 5. Lastly, section 6 wraps up the essay and prefigures the future step of our program.

2. CHALLENGES

When starting a new coding project, you are always faced with different barriers and challenges along the way. These may include using reliable data and picking the right model. We also wanted to make this project useful for the real world. Some other complications range from choosing reliable data points from the vast internet and making it easy to use for inexperienced users. Below is an overview of the different barriers we had to overcome to execute our project.

One issue we had at the beginning was searching for a reliable website source. When looking at the economy of the car market, every digit must be precise and can make a great change in how cars are exchanged. Some sources are biased while others have inaccurate information. Since every website may vary and may not be 100% accurate, we chose to use the most universal website. This reliable source comes straight from the consumer car market and uses actual car exchange prices. We agreed that extracting information from the actual market exchange is the most organic form of information we could put our hands on.

During our first stage of programming, we had trouble gathering big data for our program model. This brought us more issues as we continued further into development. Originally, we made the program to collect only five years of data from one car. But as we began testing our model, it resulted as an inaccurate and unreasonable value. So in order for the Machine Learning model to provide a more legitimate value, we must input more data into it. Ten years historical data from one car was the final decision for the model. This meant we had to revise our code's format as well as extend the lines of our input data.

Creating automation for the entire program was the biggest conflict we ran into during the process. In order for the user to be able to answer a few questions about their chosen car, we had to write a set code that somehow takes the information provided by the user and communicate with the website we chose. But we could not determine how to split apart the website's URL in the program code in order for the given information by the user to be sent. Plus, the names of the car and models must be very specific without any spelling errors since it connects with the website's URL. Developing the automation part of our program consumed the most time throughout our coding process.

3. SOLUTION

A. An Overview of the Solution

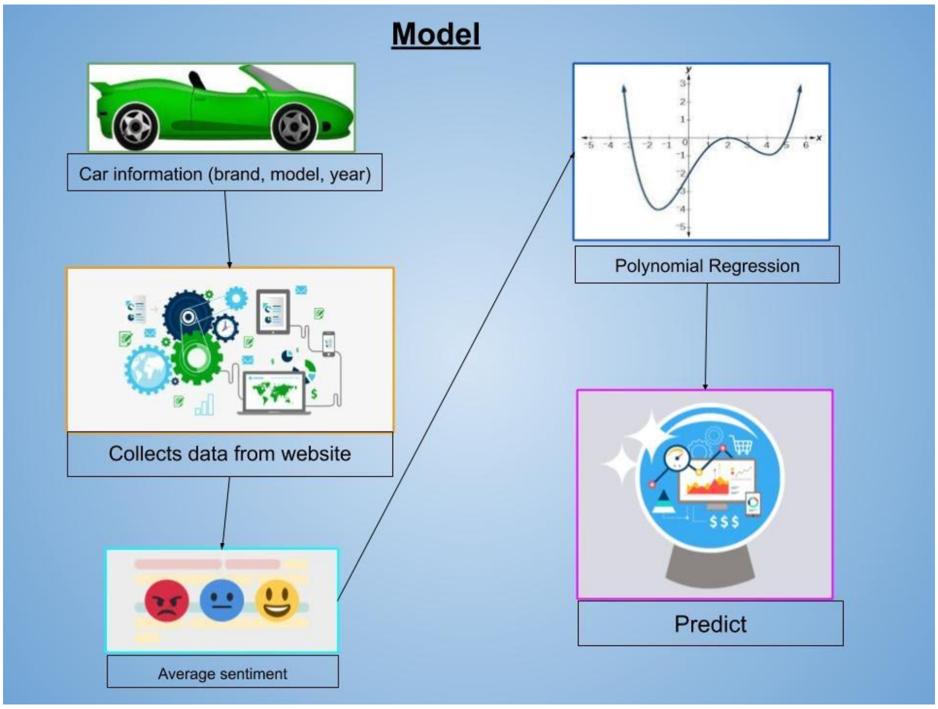
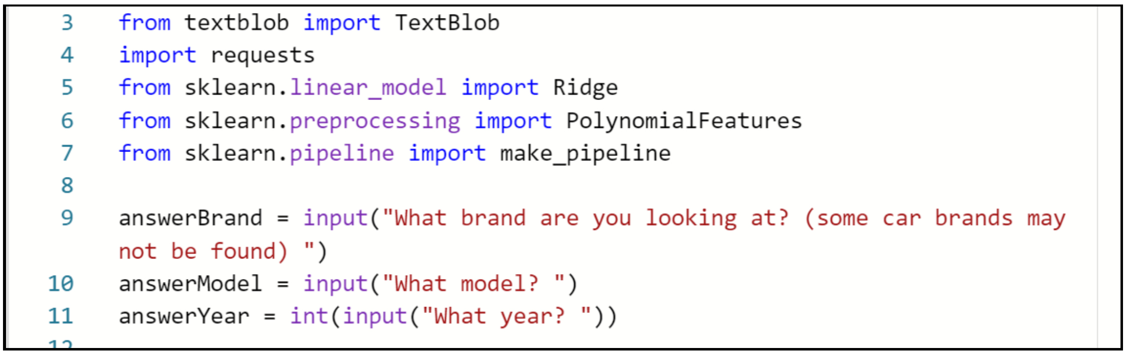

Model
Car information (brand, model, year)
Collects data from website
Average sentiment
Polynomial Regression
Predict
$$$

Figure 1. Overview of the solution

CarEnvision is able to pull the pricing of a broad variety of cars. These brands can then be specified through the model and the year it came out. Through the use of jdpower.com, we were able to pull information straight from the consumers themselves. In a sense, jdpower.com is a consumer operated system where people from the car market can sell their cars and know how much they're car is going for at the moment. After answering a short survey, CarEnvision starts running. Using TextBlob, the code was able to extract only information that was needed. We extracted thousands of reviews of different car models and implemented sentiment analysis to determine how positive the comment was. The automated system loops this code until all reviews are calculated. The grand result would be all the review's results average. This loop is then repeated for the last 5 years of the same model. Then, using Polynomial Regression we were able to predict the next year. Since the model was trained to fit cars from a variety of basic and luxurious brands, this model is able to fit most cars in the market. (Illustration shown in Figure 1).

B. Automation and Machine Learning Data

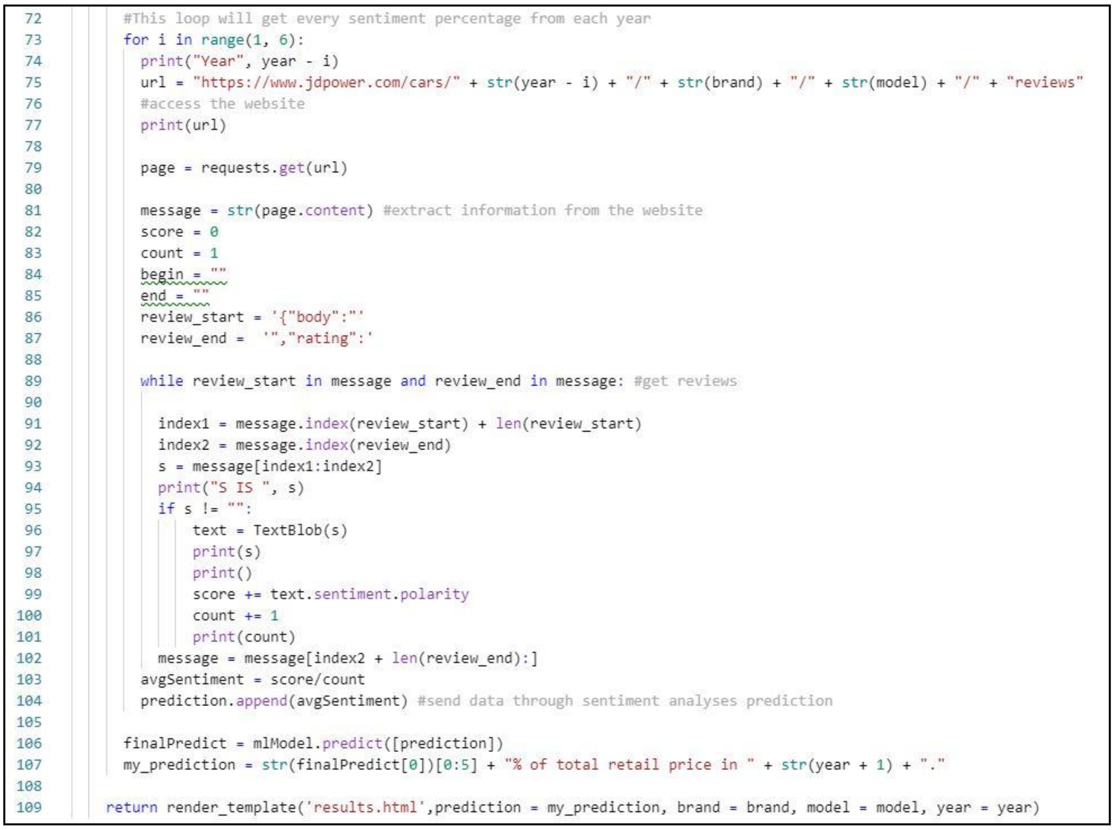
The following segments of code shows the entire program running as if a user is using it. Each component of the process is presented with a set of captured code in order to provide visual representation for the audience. The implemented code presented in Figure 2 shows the importing of textblob and different models from Machine Learning. A few lines passed, the input functions are there to collect the user's data as in the car's brand, model, and year. Each function contains different questions being asked to the user.



```
from textblob import TextBlob
import requests
from sklearn.linear_model import Ridge
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline

answerBrand = input("What brand are you looking at? (some car brands may
not be found) ")
answerModel = input("What model? ")
answerYear = int(input("What year? "))
```

Figure 2. the importing of textblob and different models from Machine Learning

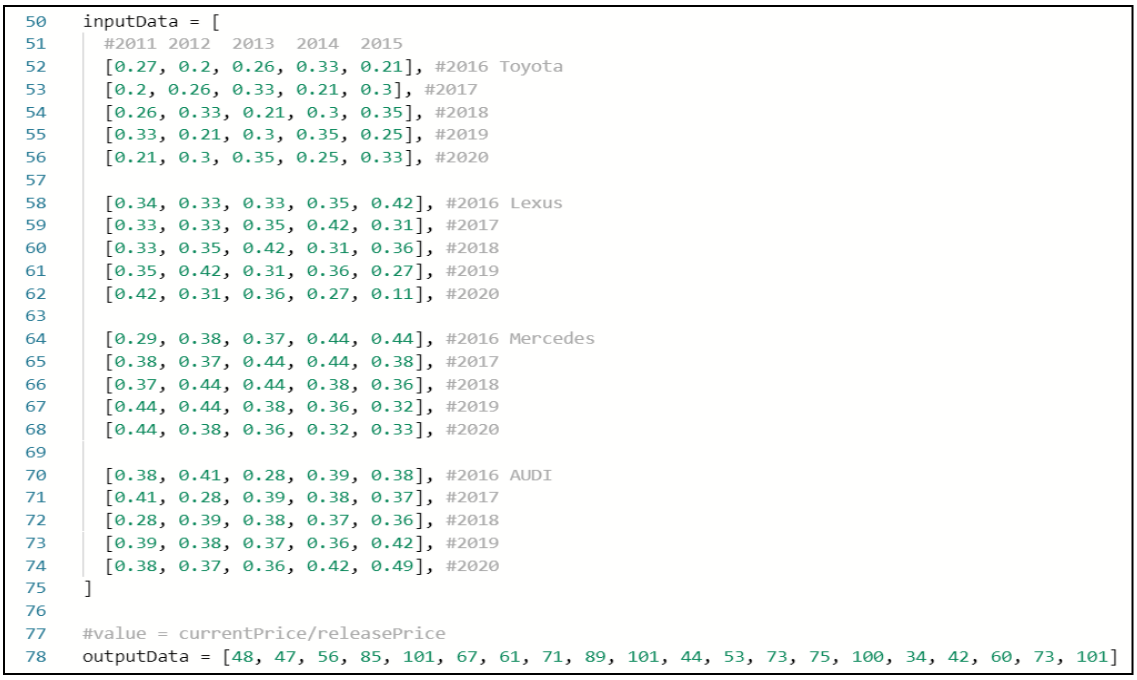
This set of code in Figure 3 is the main algorithm that operates the automation. After gathering the information from the user, each function from the previous set of code is carried to line 19 where it splits apart the URL of jdpower.com. This will take the program straight to the website and start searching for sentiments there. The loop runs for five times because it needs to gather five years of history from the car off of the website. Every year's sentiment from the car is shown to the user as an average. For example, year one represents one year before the user's chosen year and so on.



```
      #This loop will get every sentiment percentage from each year
      for i in range(1, 6):
        print("Year", year - i)
        url = "https://www.jdpower.com/cars/" + str(year - i) + "/" + str(brand) + "/" + str(model) + "/" + "reviews"
        #access the website
        print(url)

        page = requests.get(url)

        message = str(page.content) #extract information from the website
        score = 0
        count = 1
        begin = ""
        end = ""
        review_start = '{"body":"'
        review_end =  '","rating":'

        while review_start in message and review_end in message: #get reviews

          index1 = message.index(review_start) + len(review_start)
          index2 = message.index(review_end)
          s = message[index1:index2]
          print("S IS ", s)
          if s != "":
              text = TextBlob(s)
              print(s)
              print()
              score += text.sentiment.polarity
              count += 1
              print(count)
          message = message[index2 + len(review_end):]
        avgSentiment = score/count
        prediction.append(avgSentiment) #send data through sentiment analyses prediction

      finalPredict = mlModel.predict([prediction])
      my_prediction = str(finalPredict[0])[0:5] + "% of total retail price in " + str(year + 1) + "."

    return render_template('results.html',prediction = my_prediction, brand = brand, model = model, year = year)
```

Figure 3. the main algorithm that operates the automation

Machine Learning models operate only if there is enough data to train it. We chose five different example car models with five sets of different historical sentiments as shown in Figure 4. In order to train the polynomial regression model, the sentiments are used as the input data. The output data would be the current value of the vehicle compared to its first release-- found by dividing the used current price by the release price. Each car’s data was also extracted from the same website in order to maximize the accuracy of the prediction. This was the only set of data we used since we only used one Machine Learning model which was polynomial regression. We chose this model because it is capable of graphing various curves on an x and y coordinates system. Since sentiments constantly fluctuate, the graph of a polynomial would represent the set of data the best.



```
inputData = [
    #2011 2012  2013  2014  2015
    [0.27, 0.2, 0.26, 0.33, 0.21], #2016 Toyota
    [0.2, 0.26, 0.33, 0.21, 0.3], #2017
    [0.26, 0.33, 0.21, 0.3, 0.35], #2018
    [0.33, 0.21, 0.3, 0.35, 0.25], #2019
    [0.21, 0.3, 0.35, 0.25, 0.33], #2020

    [0.34, 0.33, 0.33, 0.35, 0.42], #2016 Lexus
    [0.33, 0.33, 0.35, 0.42, 0.31], #2017
    [0.33, 0.35, 0.42, 0.31, 0.36], #2018
    [0.35, 0.42, 0.31, 0.36, 0.27], #2019
    [0.42, 0.31, 0.36, 0.27, 0.11], #2020

    [0.29, 0.38, 0.37, 0.44, 0.44], #2016 Mercedes
    [0.38, 0.37, 0.44, 0.44, 0.38], #2017
    [0.37, 0.44, 0.44, 0.38, 0.36], #2018
    [0.44, 0.44, 0.38, 0.36, 0.32], #2019
    [0.44, 0.38, 0.36, 0.32, 0.33], #2020

    [0.38, 0.41, 0.28, 0.39, 0.38], #2016 AUDI
    [0.41, 0.28, 0.39, 0.38, 0.37], #2017
    [0.28, 0.39, 0.38, 0.37, 0.36], #2018
    [0.39, 0.38, 0.37, 0.36, 0.42], #2019
    [0.38, 0.37, 0.36, 0.42, 0.49], #2020
]

#value = currentPrice/releasePrice
outputData = [48, 47, 56, 85, 101, 67, 61, 71, 89, 101, 44, 53, 73, 75, 100, 34, 42, 60, 73, 101]
```

Figure 4. five sets of different historical sentiments

4. EXPERIMENT

For our experiment to test the accuracy of our model, we implemented three different models of Machine Learning to be trained by our data. Machine Learning provides several regressions for predicting different types of data. That is why we chose to compare Polynomial Regression, Linear Regression, and SVM to see which one can produce the most precise and reasonable prediction.

A. Comparison of Different Machine Learning Algorithms

In Figure 5 below, it depicts a comparison of the three models we tested. We used a 2020 Toyota Prius and a 2020 Chevrolet Camaro for the experiment. When using SVM, the prediction value is 101 for both cars which is not reasonable since it is way too high. Linear regression is not precise as well because one car got 74 percent and the other got 59 percent, and the graph is a straight line. Finally, Polynomial Regression shows the best prediction value because the result is precise and reasonable where the car value does not increase nor decrease intensely.

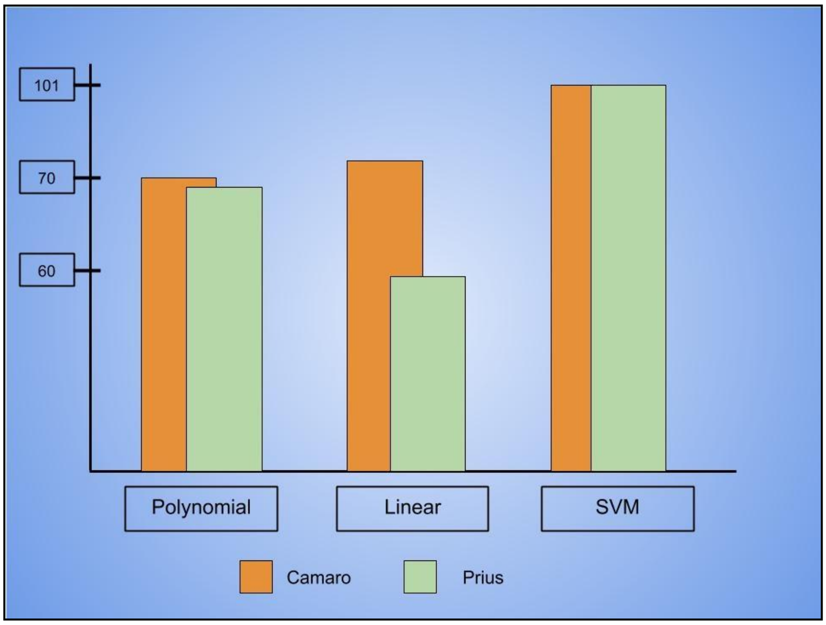


Figure 5. a comparison of the three models

5. Related Work

There are also related works and programs that perform similar functions compared to our program. For example, “Car Price Prediction using Machine Learning Techniques” by Enis Gegic, Becir Isakovic, Dino Kečo, Zerina Mašetić, Jasmin Kevrić is a program that predicts car values using data from autopijaca.ba written in PHP programming language [6]. They used a total of three models: Artificial Neural Network, Support Vector Machine and Random Forest. In contrast, CarEnvision uses only Polynomial Regression from Machine Learning and we collected data from a very universal and professional website-- jdpower.com. Cars that they predicted were also only from both Bosnia and Herzegovina, whereas our program focuses on cars within the United States. CarEnvision’s unique feature that is different from the compared example is the training data which are sentiment reviews on the cars. The experts’ opinions are what supports our prediction which is a major contrast from other projects.

Another related program that uses the same techniques as CarEnvision is “Flood Prediction Using Machine Learning Models” by Amir Mosavi [7]. His program gathers data from “rainfall and water level, measured either by ground rain gauges, or relatively new remote-sensing technologies such as satellites, multisensor systems, and/or radars” (62). Several algorithms they used were multiple linear regressions, quantile regression, and Bayesian forecasting models (34). Both of our programs used Machine Learning models but we used Polynomial regression and they included several different regression techniques. We both had different data as well as predictions; Mosavi predicted floods and we predicted car values. CarEnvision provides a much faster way to gather data since our automation just pulls the data off of a website, but Mosavi’s flood prediction had to get a piece of data daily.

Snow avalanche hazard prediction using machine learning methods by Bahram Moslem Borjia, Amir Mosavibc, Farzaneh Sajedi, Hosseinia Vijay, and P.Singhd Shahaboddin Shamshirband [8] is the last program that is similar to CarEnvision. Their avalanche prediction is based on data from “avalanche occurrence locations, meteorological factors, and terrain characteristics.” They implemented both Support Vector Machine (SVM) [12, 13, 14, 15] and Multivariate

Discriminant Analysis (MDA) as their models to predict avalanches. The only similarity our program has with theirs is we both used models from Machine Learning Algorithms. Other than that, the models fitted for the prediction was different as well as the topic of the programs. The main difference between the two programs is that our sentimental data is used to train both of our models but in different ways, while the avalanche-related data comes from three different areas.

6. CONCLUSION AND FUTURE WORK

In today's world, we are rapidly developing new technology to appease the workload of tens of millions of people around the globe, especially transportation. With modeling technology, CarEnvison makes predicting car prices seem easy. With our experiment, we were able to decide which model from Machine Learning accurately represents the data we want to input. Since the economy of the car industry changes by the minute, we needed Polynomial Regression to accommodate the different dips and ridges that may be created. Since this model represents both high and low data points, the accuracy and precision of our model greatly increased.

As with any invention, there will always be improvement as we develop. We find new ways to make the invention even better and effective. Like other programs, it may not correctly represent all brands or models. CarEnvision is limited in the fact that it cannot pull information from cars that are either not sold in the car industry or too new. In addition, the exchange of cars can include other factors and not only the opinions of the buyers, but also the appearance of the car, features, safety, or speed. As developers of CarEnvison, we solely rely on the opinions of car producers and consumers.

Due to current limitations, CarEnvison needs new ways to improve and increase its precision to help car consumers make the best decision to keep up in this rapidly increasing world of technology. Our future plans include the increase of data and inclusion of more cars and models to make CarEnvison available to more of the public.

REFERENCES

[1] Rietmann, Nele, Beatrice Hügler, and Theo Lieven. "Forecasting the trajectory of electric vehicle sales and the consequences for worldwide CO2 emissions." *Journal of Cleaner Production* 261 (2020): 121038.

[2] Ostertagová, Eva. "Modelling using polynomial regression." Procedia Engineering 48 (2012): 500-506.

[3] Peixoto, Julio L. "A property of well-formulated polynomial regression models." The American Statistician 44, no. 1 (1990): 26-30.

[4] Bradley, Ralph A., and Sushil S. Srivastava. "Correlation in polynomial regression." The American Statistician 33, no. 1 (1979): 11-14.

[5] Heiberger, Richard M., and Erich Neuwirth. "Polynomial regression." In R Through Excel, pp. 269-284. Springer, New York, NY, 2009.

[6] Gegic, Enis, Becir Isakovic, Dino Keco, Zerina Masetic, and Jasmin Kevric. "Car price prediction using machine learning techniques." TEM Journal 8, no. 1 (2019): 113.

[7] Mosavi, Amir, Pinar Ozturk, and Kwok-wing Chau. "Flood prediction using machine learning models: Literature review." Water 10, no. 11 (2018): 1536.

[8] Choubin, Bahram, Moslem Borji, Amir Mosavi, Farzaneh Sajedi-Hosseini, Vijay P. Singh, and Shahaboddin Shamshirband. "Snow avalanche hazard prediction using machine learning methods." Journal of Hydrology 577 (2019): 123929.

[9] Mair, Carolyn, Gada Kadoda, Martin Lefley, Keith Phalp, Chris Schofield, Martin Shepperd, and Steve Webster. "An investigation of machine learning based prediction systems." Journal of systems and software 53, no. 1 (2000): 23-29.

[10] Mackenzie, Adrian. "The production of prediction: What does machine learning want?." European Journal of Cultural Studies 18, no. 4-5 (2015): 429-445.

[11] Weiss, Sholom M., and Nitin Indurkhya. "Rule-based machine learning methods for functional prediction." Journal of Artificial Intelligence Research 3 (1995): 383-403.

[12] Noble, William S. "What is a support vector machine?." Nature biotechnology 24, no. 12 (2006): 1565-1567.

[13] Suthaharan, Shan. "Support vector machine." In Machine learning models and algorithms for big data classification, pp. 207-235. Springer, Boston, MA, 2016.

[14] Joachims, Thorsten. "Svmlight: Support vector machine." SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund 19, no. 4 (1999).

[15] Pisner, Derek A., and David M. Schnyer. "Support vector machine." In Machine Learning, pp. 101-121. Academic Press, 2020.

AUTHORS

My name is **Terence Chen**, one of the founders of CarEnvision. Before high school, I had zero coding experience and did not even know of this field. I started coding in Coding Minds Academy starting freshman year of high school. That was the time when the field of computer science sparked my interest. From then on, I was able to create ideas beyond my imagination through coding. In the future, I see myself going deeper into this field and formulating bigger programs as well as cooperating with many intelligent people.

My name is **Angel Chang**, one of the founders of CarEnvision. Despite having coding experience for only a year, I have been part of the Coding Minds Academy program and it has been a great accomplishment. When I first got my feet wet in the coding world, it has ever since amused me and became a hobby. Besides coding, however, I have devoted my time to playing video games, swimming, and playing water polo. I am currently in my sophomore year of high school in Rancho Cucamonga and aiming to become a future doctor.