Exploring Data Mining in Used Car Data

University of Nevada, Reno
Department of Computer Science and Engineering
CS 425/625: Project Assignment 2

Team #8:
Patrick Austin, Mile Cilic, Eric Stutzman, Guang Xu

Instructors: Sergiu Dascalu, Devrin Lee

Advisors: Stephen Williams, Lei Yang

28 October 2017

Table of Contents

Introduction

In the “Exploring Data Mining in Used Car Data” project, Team 8 hopes to use state-of-the-art data mining techniques together with an exclusive data set to discover novel and valuable information about a real-world industry.

Working in collaboration with the firm Marketing Evolution and their advisor at Marketing Evolution, Stephen Williams, the team has been granted access to a 50GB dataset documenting used car transactions in the United States over the last several years. This dataset contains attributes such as extensive demographic and financial information, as well as details about the car models being purchased and sold.

Using data mining algorithms and strategies learned from our faculty advisor Dr. Lei Yang, we hope to use a decision tree induction approach on the data set in order to train a predictive model. This model will then allow users to enter available demographic or financial details about a customer as input, and receive a recommendation of cars that customer is likely to be interested in as output. This model could be used to target personalized advertisements to a customer, or could even be packaged as an app to provide support to used car dealers on the ground.

In order to accomplish these goals, Team 8 intends to build a back end application in Python which will conduct the decision tree induction on the data, and a front end web application which can be used to query the decision tree for results. In this Software Requirements Specification document we provide extensive details about the business requirements we expect the software to serve, the technical requirements we expect this software to satisfy, the use cases we expect it to serve, and initial snapshots of the user interface we expect to implement.

Summary of Stakeholders' Interviews

Patrick is a member of our team and was interviewed to gain insight into the development team. Dr. Yang is an Assistant Professor at UNR, and was interviewed since he initially proposed this idea. Stephen is an employee of Marketing Evolution, and was interviewed as the owner of the dataset.

1. What are the success criteria for this project?

Patrick: We want to emphasize success in CS425 & CS426 classes and satisfy the other stakeholders Marketing Evolution and Dr. Yang.

Dr. Yang: The performance of the software and the accuracy of the result are most important, and how useful this result can be transferred to marketing.

Stephen: The most important thing is generating a usable prediction model that is as accurate as possible.

2. What additional features would you like see in the final product?

Patrick: I would like to see a mobile app and a suite of visualization tools.

Dr. Yang: Multi-platform access to the software would be useful.

Stephen: I would like to see an interface that allows for graphical representation of data and prediction, and the ability to perform a GIT pull with certain data that returns a value in a JSON object.

3. Who is the intended customer of this product?

Patrick: I see the customers as clients of Marketing Evolution who want to purchase analysis that uses this tool such as used car dealerships.

Dr. Yang: Used car dealerships can use the product to better advertise their cars or find the best customer.

Stephen: The product is aimed at internal employees of Marketing Evolution who deal with Carfax business, and possibly contacts at Carfax. But most likely not consumer based.

4. How can this project add value to Marketing Evolution as a company?

Patrick: The project may allow them to expand into the used car advertising area.

Dr. Yang: By trying to find strategies from the data to improve advertising, they can build a platform for a unified framework for other data.

Stephen: Marketing evolution gets value by providing a proof of concept with machine learning to optimize advertising, generate a sales demo, and have the ability to feed new data into their ROI method.

5. What kind of precision is required or desired in the product?

Patrick: I don't think we will need a high degree of precision, just a best guess.

Dr. Yang: The satisfaction of the customers using this product by increasing sales or reducing advertising costs is the most important metric.

Stephen: Precision should be as high as possible. Machine learning is new and not tested very well, so it is tough to give an exact number. I would like to see an 80% success rate or higher for a sales demo.

6. What problems are you aware of that would threaten the value of this product or service?

Patrick: Competition from other websites and big industry like Google and Facebook in targeted advertising is only growing. What we're doing in this project may be too narrow and may need to be expanded to a full suite of tools to be useful.

Dr. Yang: Privacy is a critical issue. We need to hide people's information so they cannot be identified.

Stephen: There are unknowns to machine learning and what it can do, and there may be more to the story. If we lost Carfax as a client then this would be less useful. Ultimately this is an internal product for demoing to customers.

7. Is there anything that you would like to prohibit us explicitly from doing?

Patrick: I think the team needs to focus on deadlines and attending meetings.

Dr. Yang: No.

Stephen: Avoid having a terrible time and avoid sharing the raw data.

8. What data is available for this project?

Patrick: The big dataset, other data from Blue Book, how other companies do targeted advertising, Dr. Yang, and the Data Mining textbook are all available for us to consult.

Dr. Yang: The Marketing Evolution dataset is the main source.

Stephen: The dataset and its 1.5 million rows of information on used car purchasers are our source.

9. How is the data structured?

Patrick: We have only seen a sample of the data in CSV form containing demographics like gender, age, race.

Dr. Yang: The sample is an Excel file, but I am unsure about the whole dataset.

Stephen: The data is in CSV, tabular form.

10. Has anyone else worked on this dataset, and if so can we see their results?

Patrick: Not that I know of.

Dr. Yang: No, I don't think so.

Stephen: Yes and no. I have worked on the dataset and it has been mapped. Someone is working on it now, to join with other data to paint a picture of the data.

11. What equipment and tools are available to help with testing?

Patrick: There are lots of common debugging tools and we can check theory against the textbook and other online resources.

Dr. Yang: We can look at the data from various data mining points of view to calculate results. We can split the data into 2 sets, a training set and a test set.

Stephen: We will have access to AWS, a small one all year long. Maybe you should use 3 tier for more control. Once you have a trainable model, then I can allocate time on a more powerful system for 2 to 7 days.

12. Are there any other questions I should be asking you?

Patrick: No.

Dr. Yang: I don't know.

Stephen: I don't think so.

High Level Business Requirements

#	Requirement
1.	The system shall enable Marketing Evolution to make predictions about used car sales.
2.	The system shall enable users to query for predictions given information about a possible customer.
3.	The system shall feature a user interface for both Marketing Evolution employees and potential customers to interact with.
4.	The system shall make predictions with a high degree of accuracy within a reasonable response time.
5.	The system shall enable users to make queries on multiple platforms, including via web client and via iOS/Android apps.
6.	The system shall maintain the confidentiality of proprietary Marketing Evolution data and protect the privacy of users.
7.	The system shall make economical use of Amazon Web Services for use in performing data mining computation.

Technical Requirements Specification

Functional Requirements

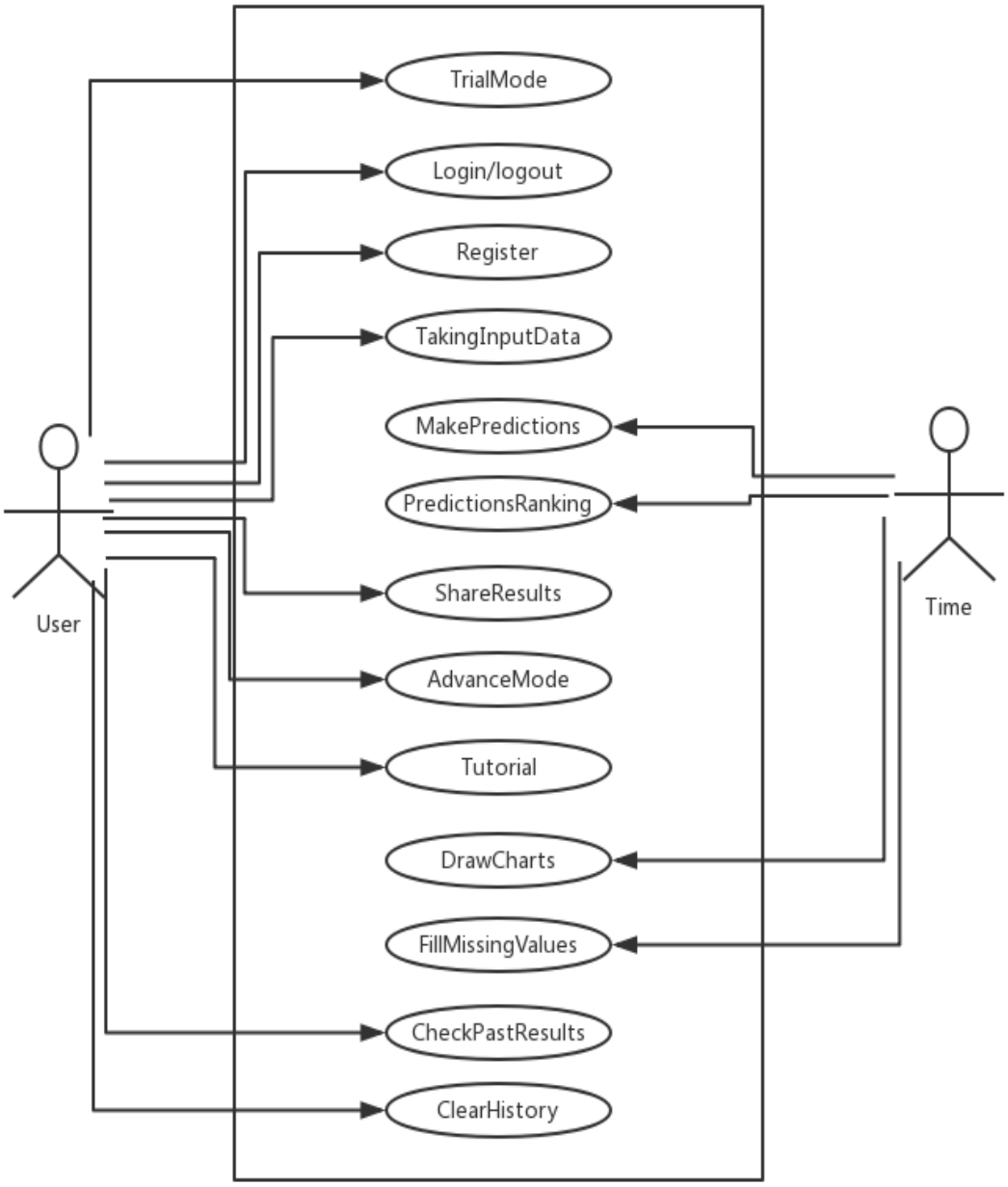
#	Requirement	Level
1.	The system shall find most profitable opportunities based on critical customer information.	1
2.	The system shall lay out the deal that customers will be most receptive to.	1
3.	The system shall deliver the highest possible ROI, while substantially reducing the cost per unit.	1
4.	The system's algorithmic data shall create powerful results that traditional methods cannot match.	1
5.	The system shall provide automotive retailers with powerful, insightful and pin-point information that identifies the highest-value customers to target.	1
6.	The system shall revise deals based on customer needs for a completely personalized approach.	2
7.	The system's algorithms shall rank customers who are most likely to trade today at the highest profit margins, allowing to focus on customers and the opportunities.	2
8.	A user shall be able to interact with the user interface that allows for graphical representation of data and prediction.	2
9.	The system shall have a multi-platform access to the software.	2
10.	The system shall provide customers with relevant, credible, and enticing information to notify them promptly when the opportunity exists for them to upgrade to a new vehicle for a similar payment.	3
11.	The system shall deliver each individual customer a new vehicle upgrade quote specific to him or her.	3

Non-functional requirements

#	Requirement
1.	The system shall hide user's information.
2.	The system shall be available to all users at all times. Downtime shall not exceed five minutes in any one day and users will be notified if it happens.
3.	Users of the system shall authenticate themselves using their personal login name.
4.	Users shall be able to use all the system functions after one hour of training. After this training, the average number of errors made by experienced users shall not exceed two per hour of system use.
5.	The system shall give response within 5 seconds upon getting a request.
6.	The mobile application's size shall not exceed more than 30MB of memory.
7.	Users will need an Android or an iOS smartphone, or a Web browser to get access to the system.
8.	Minimum system requirements include: Android 4.0 or iOS 8.0; 1GB RAM; 1.2 GHz processor; 500MB of available Hard Disk.
9.	The system, after failure, shall be able to restart within 2 minutes.
10.	The system shall provide an 80% success rate or higher for a sales demo.
11.	The system shall provide new users with a trial mode for testing purposes.

Use Case Modeling

Use Case Diagram



Detailed use cases

ID	Use Case	Description
UC01	TrialMode	New users can use the application 15 times before they register, after 15 times, their IP address will be banned from using the application unless they register and pay for the membership fee.
UC02	Login/logout	The user should have the ability to log in to the application, the application is fully open only to the members, after the tasks are finished, the user should also be able to log out of the application.
UC03	Register	After the user determines to be a member, the user can register as a member of the application, the application stores the information of the user to verify the identity of the user later.
UC04	TakingInputData	The application should be able to take inputs from the users, the inputs will be used for later calculation.
UC05	MakePredictions	After taking the inputs from the user, the application should pass this data to the server and make predictions about the outcomes based on the algorithm which is stored in the server.
UC06	PredictionsRanking	The application should provide 3 possible results and rank them based on their probabilities.
UC07	ShareResults	After the prediction is over, you can share the result with other people by clicking a single share button, you can choose from different social apps and a text will be sent to them.
UC08	AdvanceMode	The application shall have some selections for advanced users to adjust the function attributes manually to get the best prediction result. The advanced mode options should be hidden and only be available when the user selects the advanced

		mode specifically from the main menu.
UC09	Tutorial	The application shall have a tutorial mode, the tutorial mode will kick in when the user open the application for the first time, and it will take the user to go through all the essential features with guide. The tutorial mode will only be on only for the first time unless the user request to go through the tutorial one more time at the end of tutorial.
UC10	DrawCharts	The application shall take to results and draw some statistic charts based on them every time the user runs the algorithm, the charts shall be distribution charts and bar charts to show the probabilities.
UC11	FillMissingValues	The user is allowed to not fill in all the required data, as long as there are at least 3 information fields filled in by the user, the algorithm in the back end should automatically fill in any other information fields for the users with default values.
UC12	CheckPastResults	Every time the user runs a prediction, the application will store the result in the server, so the user can check his/her past prediction results in a list of view.
UC13	ClearHistory	The user can also select to delete the history of his/her results from the server, it could be all the past results or some specific ones.

Detailed Templates

	Use Case: Register
Use Case ID	UC03
Actor	User
Precondition(s)	1.The actor has installed the application on his/her device. 2.The actor is currently using the application.
Flow of events	1.The actor selects register button from the main menu. 2.The actor inputs their email and desired username and password. 3.The actor pays for any applicable membership fee. 4.The actor will get a email which has his/her receipt and order confirmation.
Postcondition(s)	1.The actor's information will be stored in the server. 2.The actor will be able to use all the features of the application without any limitations.

	Use Case: TakingInputData
Use Case ID	UC04
Actor	User
Precondition(s)	1.The actor is still within the 15 times trail limits or is logged in.
Flow of events	1. The actor will click the "run a new test" button. 2. The actor will put each piece of demographic information of a customer in the matching filed. Note the user may give at least 3 pieces of demographic information for the calculation to happen. 3. The actor will click the "run" button to start the calculation.
Postcondition(s)	1. The actor will get a prediction based on the information he/she provides.

	Use Case: ClearHistory
Use Case ID	UC13
Actor	User
Precondition(s)	1. The actor is registered and logged in. 2. The actor has run the calculation before.
Flow of events	1. The actor will click the “clear history” button. 2. The actor will then be taken to a list of previous calculations he/she ran before. 3. If the actor selects to delete a specific record. 3.1 The system will delete that specific record. 4. If the actor selects the “clear all” button. 4.1 The system will delete all the result history the actor has ever run.
Postcondition(s)	1. One or more result history of the actor will be deleted from the server.

Requirements Traceability Matrix

R/UC	UC01	UC02	UC03	UC04	UC05	UC06	UC07	UC08	UC09	UC10	UC11	UC12	UC13
FR01													
FR02													
FR03													
FR04													
FR05													
FR06													
FR07													
FR08													
FR09													
FR10													
FR11													
NFR01													
NFR02													
NFR03													
NFR04													
NFR05													
NFR06													
NFR07													
NFR08													
NFR09													
NFR10													
NFR11													

Initial Snapshots

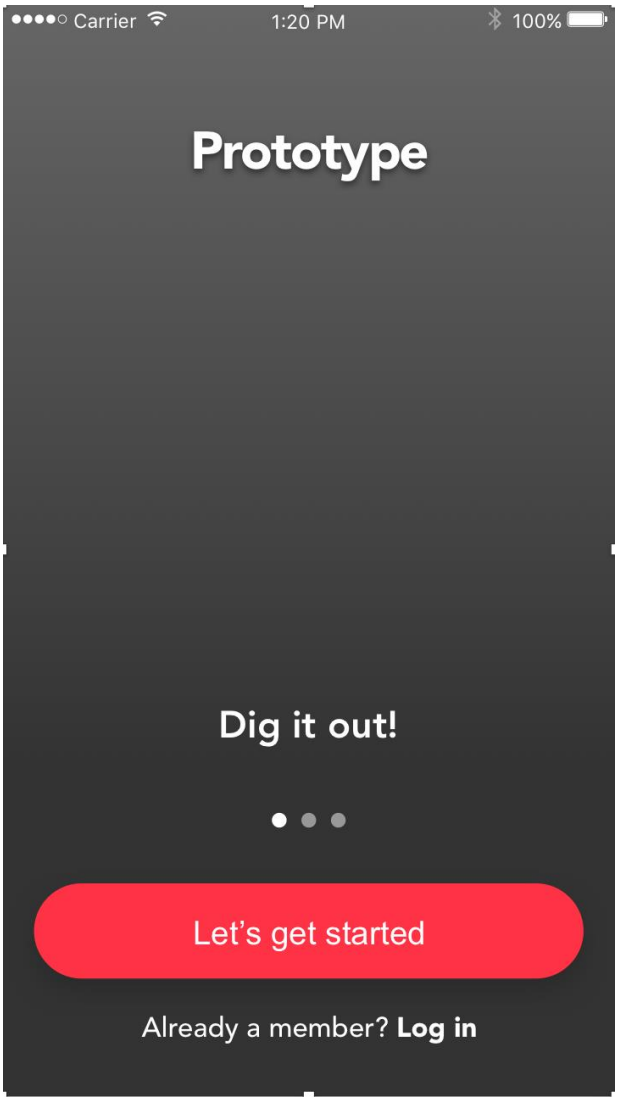


Figure 1. This is the welcome screen of our application, here you can register as an official member or log in to your account. You can slide the slogan to see more information about the app.

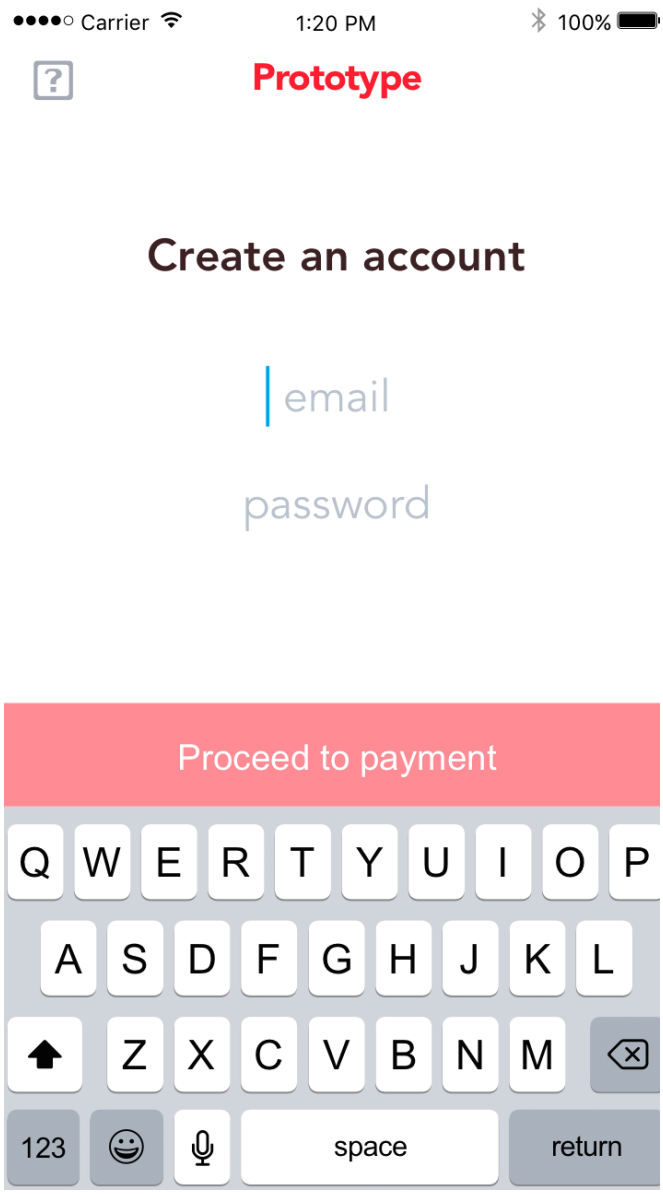


Figure 2. This is a screenshot of the customer registration procedure. Here you can put your password and email, which will be your username as well. The “proceed to payment” option will not be available until you input a valid email address and qualifying password.

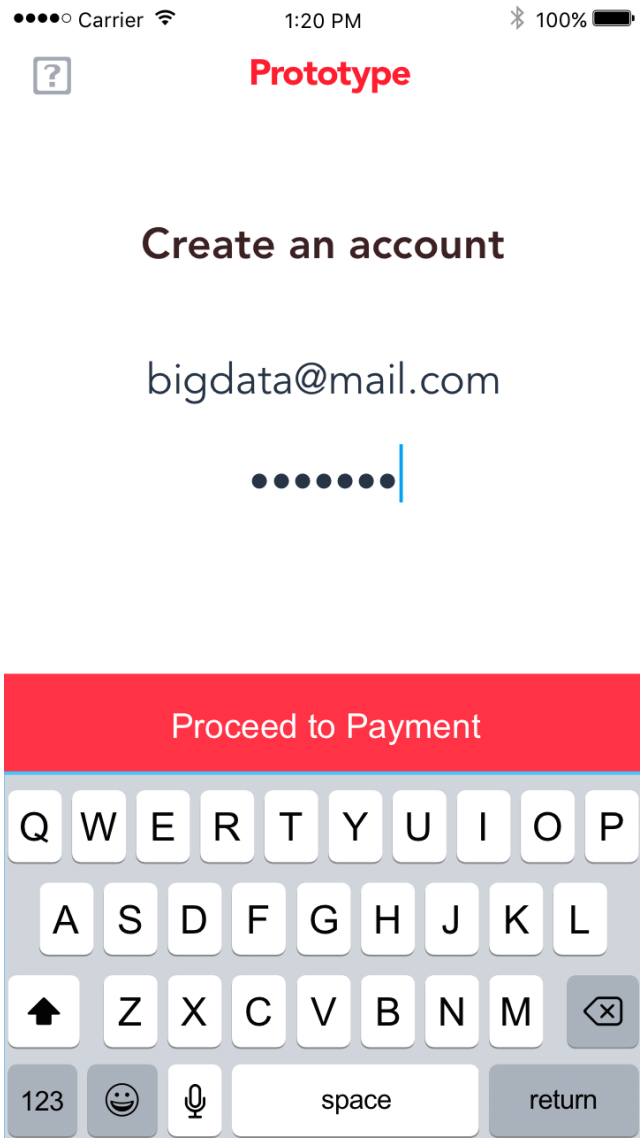


Figure 3. This is the screenshot after you input the required information, the “proceed to payment” button has become clickable.

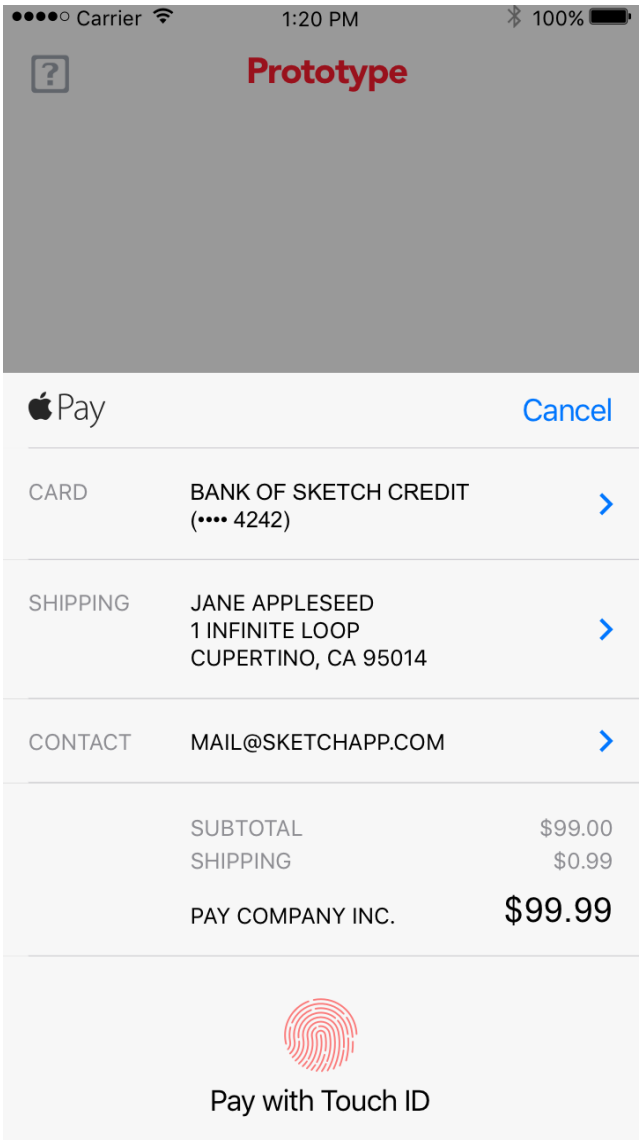


Figure 4: This is the payment page. We plan to use Apple pay to process any payments.

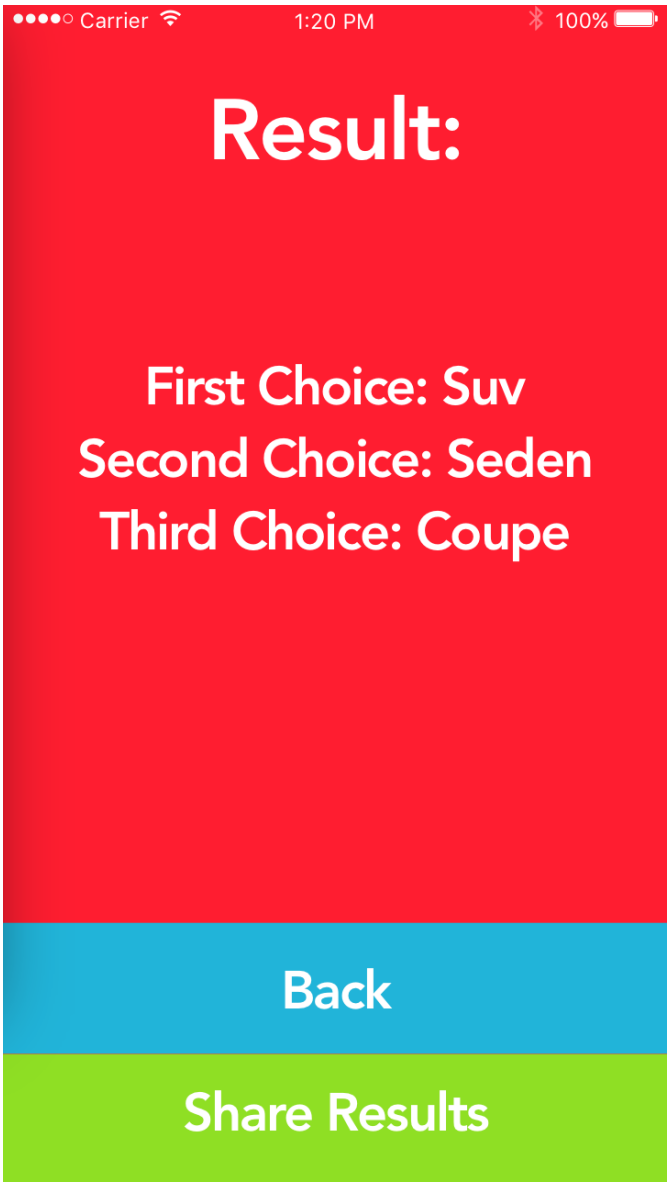


Figure 5. This is the screenshot of the result after you query the data mining algorithm. In this mock-up the customer can get a ranking of 3 types of cars that a specific person might buy. During this phase, you can click “back” to go back to the main menu or click “share results” to share the result through social applications or other applications.

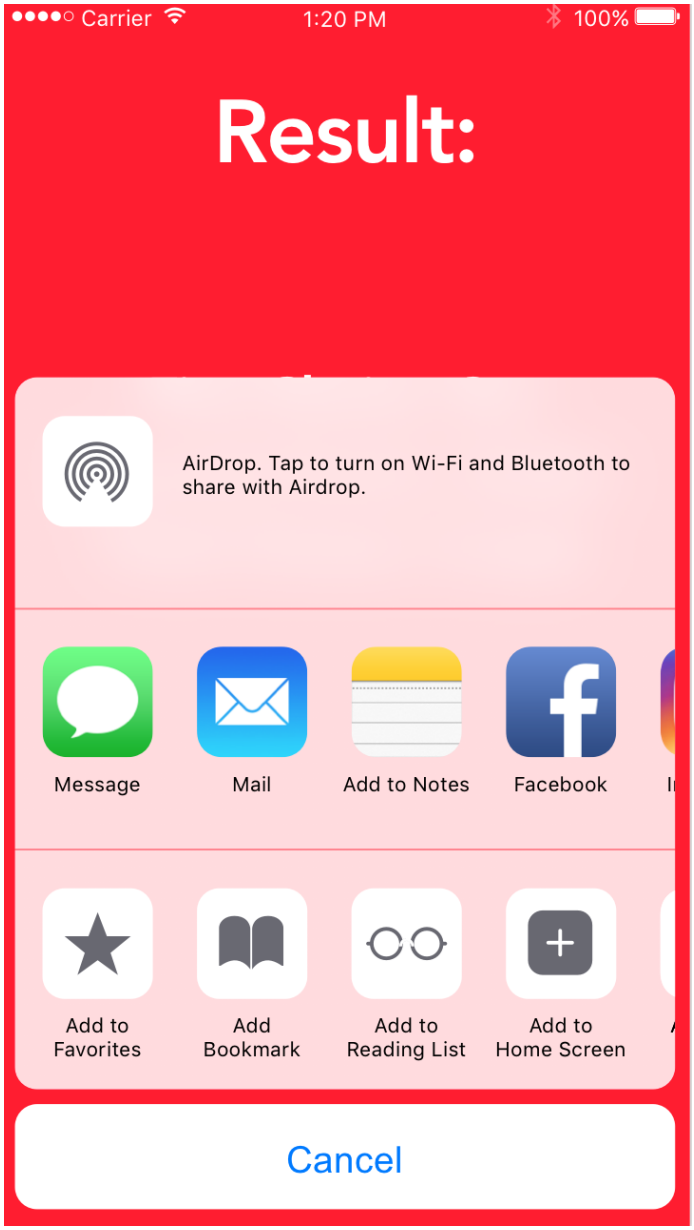


Figure 6. This image shows a share menu after you click the “share” button.

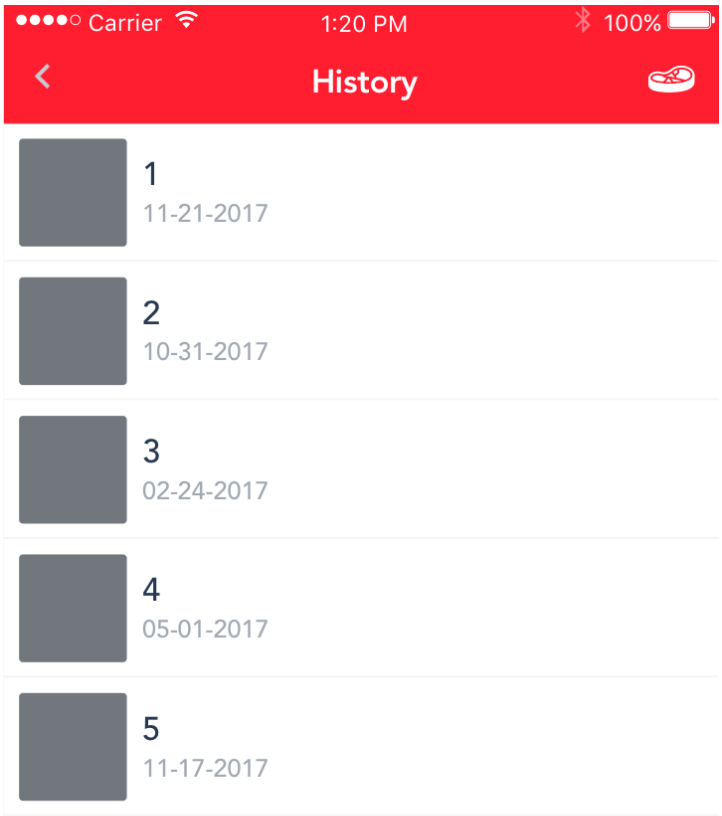


Figure 7. This is the history feature of the application, in here you can view all the past results you have run, if you click one of them, it will take you to the result page above, you can also click the back button in the top-left corner, the icon in the top-right corner is the edit button, if you click it you will be able to delete the record you choose.

Glossary

Accuracy- The ratio at which a predictive model such as a decision tree gives correct results when compared to actual outcomes.

Attribute- A property of an object or entry in a dataset, such as age or gender in a dataset of people. Corresponds to the columns in a tabular data set.

Classification- A type of data mining task where a predictive model is generated from a training data set. The model is then used to make predictions about new data.

Clustering- A type of data mining task where data entries are mapped to space, so that entries can be compared and classified in terms of their distance to one another.

Continuous variable- A type of entry in a data table whose values are represented by real numbers, such as temperature, height, or weight.

Data mining- The process of automatically discovering useful information or knowledge using large data repositories.

Decision tree- A technique for classification in which the available data is used to build a tree that branches based on the data attributes. By following this tree for a new data entry, the tree will predict some attribute of interest about that entry. For example, the tree may branch on the basis of the income of the individual being considered. For our purposes, at the end the tree will predict the car or type of car the user is likely to purchase.

Dimensionality- Refers to the number of attributes in a dataset. High degrees of dimensionality often make data mining algorithms more computationally expensive and less accurate. Several techniques to reduce the dimensionality of a data set exist.

Discrete variable- A type of entry in a data table whose possible values are finite or countably infinite. These can include binary attributes, i.e. yes or no, as well as counts and categories.

Gain ratio- A value used during decision tree induction. The gain ratio reflects whether and how much accuracy the tree would gain by splitting a node using one of the attributes.

Induction- Refers to the process by which a decision tree is built. Decision tree induction is a recursive, greedy process where splits are made so that the gain ratio is optimized.

Overfitting- A phenomenon observed in predictive data mining tasks where a model that is trained too closely on a training set loses accuracy when applied to unseen test data. Overfitting should be avoided, and pruning is one approach to doing so for a decision tree.

Performance- We use performance to refer to both the time and memory needs of our data mining algorithm as well as the success rate of our decision tree against real world data (see accuracy).

Preprocessing- Refers to the process of 'cleaning up' messy, real-world data to prepare it for data mining processing. For example, values may need to be normalized, outliers may need to be excluded, and missing values may need to be handled.

Pruning- Used to reduce the overfitting of a decision tree by removing branches of the tree that hurt the tree's accuracy when used on test data. A variety of pruning approaches exist.

Record- An individual or entry in a dataset, such as one person in a dataset of people. Corresponds to the rows in a tabular data set.

Sampling- A preprocessing approach which reduces the number of records to be considered by randomly selecting a sample of the original data set. Can be used to achieve similar results at lower computational cost, so long as the sample is representative of the original data.

Similarity- A numerical measure of the degree to which two objects are alike- for example, the distance between two objects in some space. Similarity is used in clustering approaches to data mining.

Training- The process of building a predictive model using pre-existing, available data. In this project we will train our model on the 50GB dataset, for use on new data entries.

Visualization- The display of information as a graphic or table. Visualization is used to make relationships in data easier for humans to understand and analyze.

List of References

Problem Domain Book

Tan, Pang-Ning. *Introduction to Data Mining*. Boston: Pearson, 2006. Print.

This introductory textbook on data mining is used by Dr. Yang in his CS 491 topics course on the subject. Patrick and Eric have already studied it closely in CS 491. This text contains much of the theory on data mining and the practical details of algorithms and their application on which this project is based.

Reference Articles

1. Bhargava, Neeraj, et al. "Decision tree analysis on j48 algorithm for data mining." *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering* 3.6 (2013).

This paper analyzes one decision tree induction algorithm, called J48. An open source data mining software called Weka is used to compare different approaches for creating splits on the tree in J48: either splitting using a single attribute or splitting on multiple attributes. The introductory and practical nature of this paper may be of use to us starting out.

2. Farid, Dewan Md, et al. "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks." *Expert Systems with Applications* 41.4 (2014): 1937-1946.

This paper introduces a hybrid approach to classification, using both a decision tree and another technique called a naïve Bayes classifier. This combination of techniques is shown to improve the model by various measures such as accuracy, precision, and computation time. This kind of incorporation of various data mining techniques may be useful to improve our results to the greatest extent possible.

3. Muñoz-Mas, Rafael, et al. "Comparing four methods for decision-tree induction: a case study on the invasive Iberian gudgeon (Gobio lozanoi; Doadrio and Madeira, 2004)." *Ecological Informatics* 34 (2016): 22-34.

This paper details a real-world application of classification. Several different decision tree algorithms are tested and compared for performance on a real-world case involving the presence or absence of a species of invasive fish in certain regions. As this project is all about real world application, papers like this may be of use.

Related Websites

1. “CS 490D: Introduction to Data Mining.” Retrieved from https://www.cs.purdue.edu/homes/clifton/cs490d/

This is the course website for a course on data mining offered at Purdue University. It offers freely available slides, projects, and problem solutions. As this course is based on a different data mining textbook than the one we are using, it may be valuable as a resource to refer back to.

2. “scikit-learn: machine learning in Python - 0.19.1 documentation.” Retrieved from http://scikit-learn.org/stable/

This is the website for the scikit-learn library in Python, an open source library supporting a variety of statistics and data mining applications. As this is a library we are likely to use in our project, the documentation and example code available here is likely to be referred back to frequently.

3. “Mining of Massive Datasets.” Retrieved from http://www.mmds.org/

This website offers another alternative textbook on data mining, along with an accompanying MOOC containing slides and video lectures. The MOOC is associated with Stanford University. This website provides us another resource for cross-referencing and comparison on any points where the Tan textbook and our other available resources prove insufficient.

Contributions of Team Members

Team Member	Task/Contribution	Time Worked
Patrick	Introduction	30 minutes
	High Level Business Requirements	45 minutes
	Glossary	1 Hour
	List of References	1 Hour
Eric	Summary of Stakeholders' Interviews	3 hours
Guang	Use Case Modeling	1.5 hours
	Initial Snapshots	2.5 hours
Mile	Technical Requirements Specification	1.5 hours
	Requirement Traceability Matrix	30 minutes

In addition to time spent working on individual sections, the team conducted several meetings to coordinate, discuss, and ensure the quality of the document. These meetings lasted approximately 4 hours total.