Leveraging Dwell Time Models to Create Dynamic Vehicle Services

Carnegie Mellon University, Tepper School of Business

Masters of Science in Business Analytics Program, 2022

Chatchawan Lakkhananukun
Prakhyat Pola
Rebecca Stevens
Jocelyn Wang

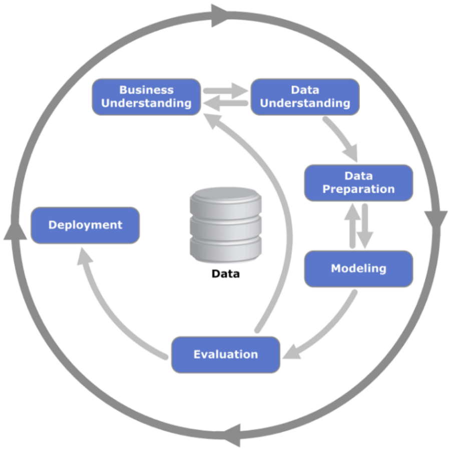
Table of Contents

Project Overview

Our team had the opportunity to partner with 99P Labs, a research group focused on coming up with innovative features, concepts, services and designs to remain on the cutting edge of the mobility industry and potentially transform the landscape of transportation. In a recent related project, a team of Berkeley students built a model to predict the dwell time and location of vehicles based on car sensor and infrastructure data from a leading automobile manufacturer. In the problem we've been given, the main question we want to address is: If we are able to predict the dwell time and location of our vehicles, what business opportunities can we leverage with this information?

Project Framework

We used the CRISP-DM (Cross Industry Standard Process for Data Mining) framework as a guide as we worked through the project. Within this framework are six consecutive processes that we will discuss in the following sections (see figure below): Business Understanding, Data Understanding, Data Preparation, Modeling, Validation and Deployment.



CRISP-DM Framework

Business Understanding

With the Telematics dataset, we want to explore the use of predictive models to add business value. As a starting point, we brainstormed business ideas that could leverage location and dwell time predictive models, which will be covered in more detail in the next section. One idea in particular is to provide vehicle maintenance services, which we believe aligns well with the 99P Labs' existing capability to offer automotive services through car dealerships. This would offer a convenient way for customers to receive routine vehicle maintenance both at home and in densely populated areas (e.g. work, grocery store, shopping center, etc.). Not only can car sensor data provide real-time feedback to the company, but dwell time derived from features in the dataset and location can also provide an estimated timeframe and optimal location for the service provider. Another business idea we explored was a peer-to-peer car lending service. With the increasing rate of adoption in car sharing services like Turo, it is possible for 99P Labs to take part in the market. One possibility is that the dwell time prediction can determine the time when cars are idle and car owners can place it on a platform owned by 99P Labs. This is an opportunity for car owners to earn extra money and let their assets work while they don't use their cars. The location prediction can provide information about the best designated area for lenders and renters to drop/return the cars. 99PLabs and its customers can each earn a portion of the renting fee.

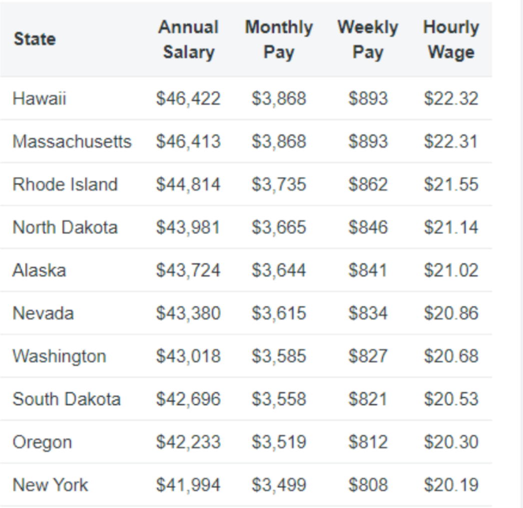
While we had the previously-built predictive models at our disposal, these dwell time and location models were geofenced to the Ohio region and were built without a focus on deriving feature importance and model interpretability. Without these kinds of model insights, it is difficult to provide clear and actionable recommendations that are valuable to the business. Therefore, another objective of ours is to identify important features that contribute to location and dwell time results. Lastly, we want to develop a framework that 99P Labs can use to validate models and determine production readiness.

Vehicle Maintenance Service

Background Research and Understanding the Competition

Before we develop a strategy, we need to first understand the services already available on the market by direct competitors. An OEM currently offers a wide array of at-home maintenance services and uses remote diagnostics to pre-diagnose repairs. The services and vehicle overview are all available to their customers through an integrated app, which allows the OEM to send service reminders, product recommendations, and other useful information. While the car is still under warranty, there is no charge for routine repairs. Outside the warranty, the OEM charges a labor rate of $150/hr on average.

There are other convenient services. Through a website, customers can schedule general maintenance for nearly every part in their car. They offer services for many car brands and makes/models. When you select a service, additional recommended services are offered that might also be required, along with the recommended service intervals. They then generate an estimated service time and allow customers to schedule the repair, at which point they pair customers with mechanics based on proximity. This service is very versatile, but requires a lot of user input. We can improve on this model by automating services, providing subscription services, and fine-tuning product recommendations based on consumer and car history. Prediction models can reduce the number of required interactions from customers, provide push strategies, and help service providers to become more proactive in reaching out to customers.



State	Annual Salary	Monthly Pay	Weekly Pay	Hourly Wage
Hawaii	$46,422	$3,868	$893	$22.32
Massachusetts	$46,413	$3,868	$893	$22.31
Rhode Island	$44,814	$3,735	$862	$21.55
North Dakota	$43,981	$3,665	$846	$21.14
Alaska	$43,724	$3,644	$841	$21.02
Nevada	$43,380	$3,615	$834	$20.86
Washington	$43,018	$3,585	$827	$20.68
South Dakota	$42,696	$3,558	$821	$20.53
Oregon	$42,233	$3,519	$812	$20.30
New York	$41,994	$3,499	$808	$20.19

Top ten average auto mechanic salaries, by state

Driven away from dealer-owned repair shops due to high prices, consumers often turn to local repair shops for routine repairs. Most independent auto repair shops charge a labor rate of $80–90 per hour as compared to the dealership average of $85–125 per hour.

The figure above shows the top ten average auto mechanic salaries by state. We will need to make up for this price difference by refining our product recommendations, add-ons, and capitalizing on the convenience of the services we offer.

Proposed Services and Benefits to the Manufacturer and its Customers

The package we envision as a solution will involve a slew of vehicle maintenance services. The idea is to add on to the existing suite of services offered through the Car Connectivity App, with features such as vehicle notifications and dashboards. We can leverage the predicted dwell times generated by our models to offer personalized services when our customers are home. We plan on basing the logistics of the service for vans and auto mechanics on an OEM's at-home maintenance model. There are several reasons why this offer will appeal to our customers, as well as why it will benefit 99P Labs itself:

- Consumers need not worry about waiting at repair shops or needing to be picked up if the repairs are extensive. Together with the predicted dwell time and location, the manufacturer can offer services that fit into the estimated service timeframe and location that is optimal for maintenance service. We hypothesize that customers will appreciate zero wait times for car services.
- 99P Labs can leverage models for a push marketing strategy to attract customers and reach a larger customer base.
- Using car sensor data and service history, the car manufacturer can plan pre-diagnosed repairs and order parts in advance; this will allow consumers to avoid lengthy lead times on required parts.
- This service is great for increasing brand-loyalty and discourages customers from having repairs done at third-party garages. This in turn will allow the manufacturer to promote additional in-store products for purchase when the customer schedules a servicing appointment.
- By maintaining a more detailed service history on the vehicle, 99P Labs can offer more advanced services that would require customers to bring their car to a dealership — this can include customer drop-offs and scheduling reminders.
- The team recommends that 99P Labs invest in developing an app to integrate available services and general info for consumer reference or add onto the existing Car Connectivity App. This platform can also be used for season promotions on tires and add-on equipment.
- By offering a customer-focused service experience, 99P Labs can open up numerous opportunities for additional product placement and advertising. This leads into the pricing aspect of the business plan.

Pricing Plan and Business Opportunities

There are several hypotheses when it comes to charging for these maintenance services. The most basic plan involves charging customers on a case-by-case basis. We can recommend a service through the Car Connectivity App, offer price overviews, and allow customers to schedule appointments. We can then offer discounts based on customer driving history and a simple points and rewards system based on previous purchases. It's possible to partner with insurance firms as well to exchange data on driving patterns as observed by the car sensors, allowing them to adjust rates accordingly. Safe driving can be rewarded with discounts on routine servicing.

We also have the offering hypothesis involving subscription services. This can be for seasonal maintenance, routine checkups, and oil and fuel replacement. Customers can receive discounted rates by subscribing to yearly packages with different maintenance scopes as opposed to scheduling one-time visits. We can use the app to send reminders for upcoming appointments and status updates. By using our dwell time data, we can plan these services well in advance, allowing 99P Labs to optimize the use of their manpower and technician availability. It will also allow service centers to maintain more accurate margins on their part orders, since they will have a detailed record of previous and upcoming maintenance activities.

Finally, we can promote additional in-house products through the app, around the time of the scheduled visits as add-ons. This discourages customers from finding cheaper alternatives at third parties and gives the customer a more user-friendly experience by providing a one-stop shop for everything they need.

Peer-to-Peer Vehicle Lending Service

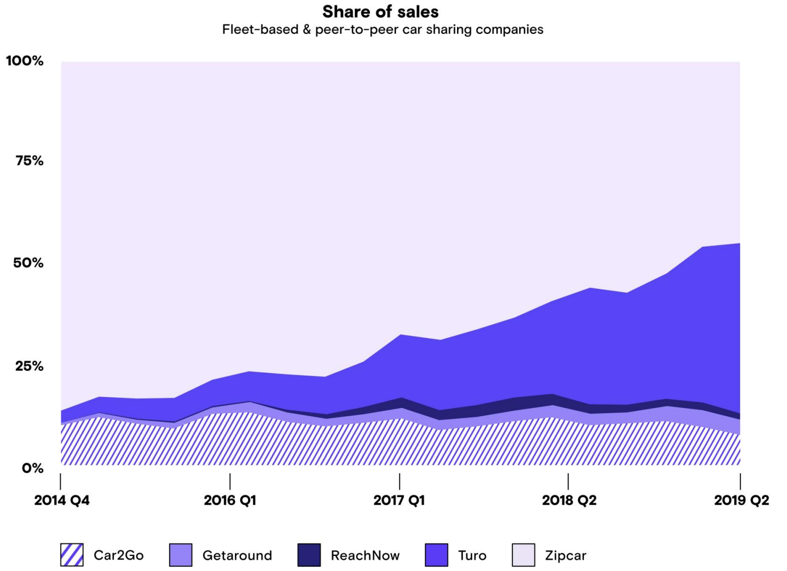
Background and Research

For our second business plan, we intend to develop a peer-to-peer car lending service bolstered by our predictive models for customers' dwell times and locations. This service is similar to Turo, a popular decentralized car rental platform. At the beginning of 2022, Turo's network encompassed roughly 450,000 listings across the U.S. and Canada, with more than 14 million users in 5,500 cities. Unlike traditional car rental services, Turo allows renters to book vehicles directly from private car owners. Some benefits of this model are:

- Turo has very low overhead relative to traditional car rental services.
- Turo is able to scale their percentage taken from each transaction based on the insurance options their customers choose (minimum, basic, or premium coverage).
- Customers are not limited in the make/model they choose: they are able to rent any car available on the app or website.
- Cars offered are generally in better condition and better-maintained than traditional fleet-based car rental agencies.

Users can rent cars for travel, as well as to test drive cars they are interested in purchasing. The car owners earn income from each of these transactions. Turo's main source of revenue however, is through the percentage of the insurance renters pay, anywhere from 10-40% of the cost of coverage per transaction. Insurance is required for everyone on the platform including the hosts; in the US, Turo offers hosts up to $750,000 in third-party liability insurance through Liberty Mutual, with five different plans to choose from.

Peer-to-peer car lending platforms are disrupting the traditional fleet-based car rental industry. The most recent example of this is Car2Go, which pulled out of several markets in 2020 due to overinvesting in their fleet size and underestimating the number of resources required to continue day to day operations. Enterprise Car Share, ReachNow, and LimePod have all shut down operations in most of their cities due to the amount of competition present and financial burdens due to: rising insurance costs, parking premiums, maintenance costs, and telematics devices. Peer to peer car sharing platforms like Turo are able to avoid these operational costs by placing the onus on private car owners rather than having to manage their own large fleet. Private car owners already pay for their parking and vehicle maintenance, and Turo even allows hosts to opt out of Turo's insurance programs in favor of their own; this means Turo's income is essentially all net gain. In addition, Turo can avoid the problem of fleet distribution across their cities since peer-to-peer lending platforms are naturally more diverse and available in more locations. The graph below shows just how large a disruption peer-to-peer lending services like Turo are causing in the car rental industry:



*Source: Second Measure

Application of Peer-to-Peer Lending Models for Honda Vehicles

We believe Honda and 99P labs can take advantage of this growing trend by introducing their own peer-to-peer lending service, targeted at current Honda customers. The number of peer-to-peer car sharing vehicles globally is expected to reach approximately 990,000 vehicles by 2025 according to Accenture's recent research; the peer-to-peer market is expected to grow to $21 billion by 2030 in China, US, and Germany alone. Instead of taking a percentage of the insurance fees from each transaction, we can instead take a rental commission from every transaction or introduce mileage-based fees, alleviating some of the insurance restrictions placed by companies like Turo. Honda currently manufactures cars in the US, Canada, Mexico, Brazil, China, Thailand, India, and Vietnam. It's safe to say there are Honda distributed nearly worldwide, presenting a huge potential market for a peer-to-peer lending service. A resilient business model, combined with diverse market regions and a data-driven marketing strategy has great potential for success; we can even add this functionality to existing Honda apps such as HondaLink, reducing the need for developing and maintaining a brand new platform for this service. This service will be especially popular in crowded cities, where many people may currently prefer public transportation over the hassle of owning a car and paying premium parking prices; daily commuters especially will be attracted to the flexibility introduced by being able to have a car readily available as opposed to waiting for trains or buses. Finding hosts should be simple considering the numerous benefits hosts stand to gain, as the revenue from offering their car for rent can help pay off loans, insurance, and maintenance costs. Studies show that most private vehicles sit idle more than 90% of the time; there is very little opportunity cost involved with lending a car while a host is at work or working from home. Because the peer-to-peer car rental space is still not as prolific as the house sharing industry (think Airbnb), this is the perfect time for Honda to invest in such a service; in addition, Honda's reputation as a quality car manufacturer and as a long-standing business will help alleviate concerns renters may have about the safety of their transaction. This model can also be extended to bike sharing as well, albeit a more niche market. The main challenges we anticipate facing as we move to implement this model are ensuring there is a large enough market for this type of service, and integrating our machine learning models into our business plan. This will be the main focus of our second survey.

Data Understanding

Telematics Dataset via the 99P Labs Portal

Several datasets are available via the 99P Labs data portal. Our focus in this project is on the Telematics dataset, which consists of over 25 million rows and 287 features. Car sensor data was collected in real-time — when an event happens, the corresponding sensors will output data and a row gets inserted into the dataset. Not all sensors will generate a signal every time, therefore we are left with a fairly sparse dataset. The challenge for us was to find clever ways to aggregate each feature variable to create insightful predictive models. Due to the long download time, we retrieved a subset of 25 million rows from the entire dataset. Within this dataset, we found that:

- There are 5129 unique vehicles.
- Many columns comprise almost all null values or contain highly imbalanced classes for categorical data.
- There are missing consecutive sequences for each car. This may affect location prediction accuracy.
- Many data tables (groups of sensor data) will not have a high impact on the predictive models (e.g. Media, Satellite). We decided to exclude these tables.
- Some features such as average temperature may be powerful in determining dwell time, but have too many missing values. An alternative approach to this is to get historical temperature data.
- There are other interesting features that we can explore as a target variable, depending on the business use case. For example, an aggregated feature from the Diagnostic table can be used for car maintenance services.

The data represents 2 weeks of car sensor events. We know that any models built based on this limited time frame will not generalize well to new data. Driving behavior may change, points of interest may shift, or seasonality may not be captured in the data. Nevertheless, the size of the data itself proves to be a challenge for our team because we had limited experience dealing with big data.

Survey Data

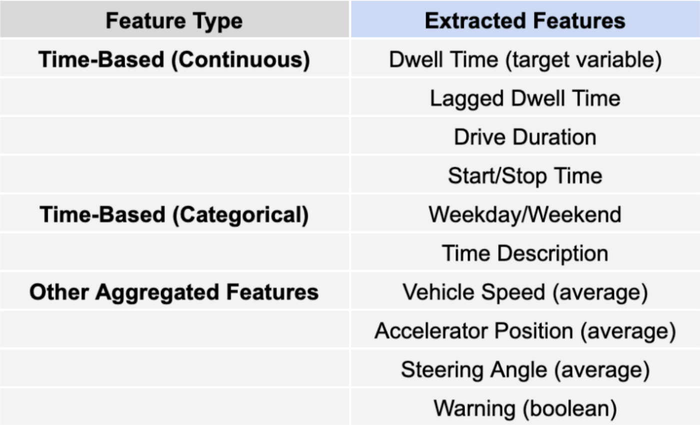
We lacked some confidence in the available datasets to fully support our business use cases (car maintenance service and peer-to-peer car lending), so we conducted a survey to validate our ideas and understand customers' patterns. The collected survey results can also be used for marketing purposes. The details of this survey are covered in a later section.

Data Preparation

Telematics Dataset

Due to the large size of the data and our limited disk storage (RAM), it is almost impossible to load the Telematics dataset using the widely-used Python library, Pandas. As an alternative, we explored using Dask, a library based on Pandas, to work with a dataset this size. In the end, we leveraged PySpark, a Python interface to the well-known Apache Spark, to work with the data as made convenient with the computing resources provided by 99P Labs. PySpark is suitable for big datasets due to its ability to distribute the data processing to a cluster of computers. Furthermore, its lazy loading feature, which loads the data partially, allows us to process the data without having to load the entire dataset into memory.

We focused on building a model for dwell time prediction and to identify important features. We grouped the data by vehicle ID and Sequence to represent individual trips, which allows us to cut down the number of rows considerably in the dataset. On average, there are 43 trips per car and with 5129 cars, so we can expect around 220k remaining rows. We extracted time-based features, location features and other aggregated features based on assumptions that they will give high predictive power.



Feature Type	Extracted Features
Time-Based (Continuous)	Dwell Time (target variable)
	Lagged Dwell Time
	Drive Duration
	Start/Stop Time
Time-Based (Categorical)	Weekday/Weekend
	Time Description
Other Aggregated Features	Vehicle Speed (average)
	Accelerator Position (average)
	Steering Angle (average)
	Warning (boolean)

Cleaned Model Features

Location Clustering

Directly feeding latitude and longitude into predictive models may not provide meaningful input to predictive models. We followed the Berkeley team's approach by clustering locations using unsupervised learning, but with a significant number of additional features.

The most common algorithm to start with is K-means, but selecting the number of clusters (K) would be difficult without the proper domain knowledge to make an educated guess. K-means is a distanced-based clustering method, and the measure will converge to a constant value between any given example as the number of features increases.

We determined that a better approach would be to use DBScan, which clusters data points based on data points' density and is insensitive to outliers. Another advantage to this method is that we don't have to select a number of clusters. However, two hyperparameters must be tuned:

- **epsilon:** the maximum distance between two samples for one to be considered in the neighborhood of the other
- **number of minimum samples:** the number of samples (or total weight) in a neighborhood for a point to be considered as a core point

Careful selection of the hyperparameters is therefore important for the resulting clusters. We included dwell time, drive duration, weekend, description of time, longitude, and latitude as inputs to the clustering algorithm. The output from DBScan was 612 clusters. One downside to this approach is that interpreting the clusters is difficult. We ultimately appended the clusters to the main dataset as an additional feature, learned through this unsupervised learning exercise.

Location Prediction

The location clusters reported in the previous section may not be the optimal model if we want pure location information. The clustering model contains information about other features that are affecting the cluster's output. Hence, with a similar approach, the location prediction model involves a series of clustering and regression, but only with location features, namely the longitude and latitude. First, the last longitude and latitudes of each sequence (trip) were extracted from the dataset. Similar to the first step in forming a cluster for dwell time prediction, we applied the DBScan algorithm, a type of clustering method, to the longitude and latitude to form a set of clusters. The metric used to measure distance was *Haversine distance*, which measures the angular distance of a sphere surface and is suitable for measuring distance for longitude and latitude. With the centroid of each cluster, we fitted an XGBoost regression model to predict the cluster. Since there is an issue of information leakage of location with the dwell time model, we did not specifically use the location prediction in our business model.

However, because both of our business models require location information to be successful, we propose a similar approach to the method mentioned here and extract key information such as zip code, city from the predicted longitude and latitude. A simple starting point would be to use the centroid of each of the clusters and use location API to extract this information.

One important thing to note here is that in order to predict both dwell time and location and to be useful in our business models, information leakage must be taken care of. Since the dwell time prediction uses the location clusters, additional location predictive models will be rendered useless. With the limited time, we did not fully explore the possibility of eliminating the location features in order to predict the dwell time, and thus would be one of our recommendations for 99P Labs to continue looking into this.

Survey Validation

Vehicle Maintenance Use Case

We decided to conduct consumer research through a survey medium and ask basic questions regarding vehicle and service/repair patterns. We wanted to use this survey as a way to validate some of the trends we were seeing in the Telematics dataset, as well as some of the assumptions we were making for the business plan. We distributed the survey among our own networks, both professional and personal. We had a total of 190 people take the survey, though only 162 completed the survey in its entirety. Out of the 162 respondents, 154 said they had at least one vehicle in their household. This pool of 154 respondents became our sample for analysis. We discuss some notable Tableau analysis results below.

Respondent Demographics

This sample pool was fairly evenly distributed throughout the different demographics, though it was more heavily weighted in the age category of 25–34 years old (58% of respondents). This is likely due to the fact that our team members are part of this age bracket and, therefore, the majority of our networks are as well. Though we would have liked to see a smoother distribution between all ages, we ultimately were satisfied with the results as we had such a successful response rate overall, and this age bracket is also our biggest focus for our target market. In addition, our responses were mostly distributed between those who reside in suburban or urban areas (94%); very few respondents said that they reside in rural areas. This was favorable for us, as we want to target those who reside in more heavily populated areas.

Vehicle and Maintenance Behavior

As we assumed, the majority of repairs people have done on their vehicles are considered routine maintenance (i.e. oil change, tire rotations, vehicle inspections, etc.). For this routine maintenance, respondents are fairly split between whether they bring their vehicle to an independent garage or their franchise dealership. This is definitely something that we would need to consider in our business plan development and could be considered an obstacle, as we would need to consider ways to entice people to choose their franchise dealership for maintenance instead — in this case, our partnering car manufacturer.

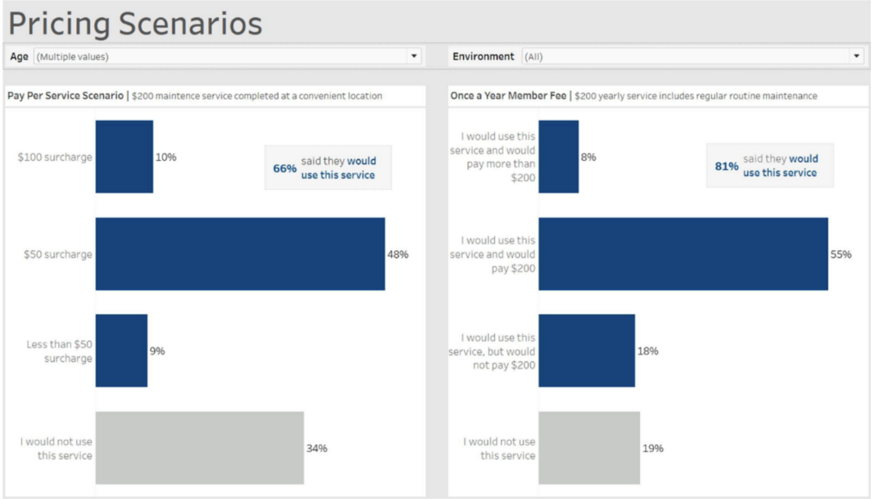
Driving Patterns

We asked the respondents questions about the driving patterns, for example where are the most common places they travel to, as well as how far they travel to get there and how long they dwell at these locations. These questions served as a great way to validate some of the trends we saw in the Telematics datasets. For example, nearly 70% of people are making primarily short distance trips (0–19 mile radius). Some of the most commonly visited locations include the grocery store, work, restaurants, and other stores/shopping centers which all were chosen about 50% or more of the time. In addition, we typically saw that when people travel longer distances, they are much more likely to dwell for a longer period of time. This suggests that distance traveled and dwell time are positively correlated.

Pricing Scenarios

The last important topic of questions we asked in the survey was centered around a couple of different pricing scenarios. These questions were posed in order to get a feel for what kinds of services respondents were interested in, and what type of pay schedules they were willing to subscribe to.

The first question posed the hypothetical scenario of getting maintenance done at a vehicle owner's local garage, consisting of 3 hours of labor and costing $200. We wanted to know if they would be willing to pay a surcharge in order to have the technician come to their home or other convenient location such as work, to complete the maintenance there. If they were interested in this service, we asked further how much of a surcharge would they be willing to pay. Out of the 154 respondents, a total of 59% said they would use this service and 44% said they would pay a $50 surcharge.



Survey results for pricing scenario questions.
Scenario #1: pay-per-service surcharge (left);
Scenario #2: once a year membership fee for routine maintenance

The second scenario asked if they would be willing to pay a yearly membership fee of $200 that would cover regular routine maintenance such as oil changes, tire rotations, vehicle inspections, etc. Out of the 154 respondents, 76% said they would use this service and 50% said they would pay $200. What was interesting to find was that if we filtered the results to include only those aged 18–44 years old, excluding those not in our target market, the results improved (shown in the image). When looking at the first scenario, (on the left), that offered a pay-per-service surcharge example, those that were interested in the service increased to 66%. The second scenario (on the right) that offered a once a year membership fee for routine maintenance increased to 81% of people interested in the service. At the end of the day we learned that people are much more likely to subscribe to a once a year fee over a pay-per-service surcharge payment schedule.

Peer-to-Peer Vehicle Lending Use Case

In this section, we present the survey data analysis and how it is fitted as part of our analysis of the peer-to-peer vehicle lending business case. We used Dscout to send out this second survey to 200 people. In addition to similar demographic questions as the first survey, the focus this time around was to gain insight into whether or not people would consider taking part in the peer-to-peer vehicle lending service. Ultimately, 72% of respondents said they would consider lending their personal vehicle and 91% said they would consider renting someone else's personal vehicle. Those who would be willing to rent out their vehicle said they would likely do this in the case that they have an extra vehicle that sits in their garage, they are taking a trip and are not home, or they are home for the weekend. For those interested in renting someone else's vehicle, it would most likely be when they are taking a local day trip, or they're moving and/or need a specialized vehicle for transporting items. Both renters and lenders agreed the preferred location for pick up/drop off would be a public location designated by the platform. In addition, we learned that, should a renter return the vehicle late, the lender would take an extra fee from the renter as compensation. In this situation, 90% answered in this manner and said they would not delete their account should this happen.

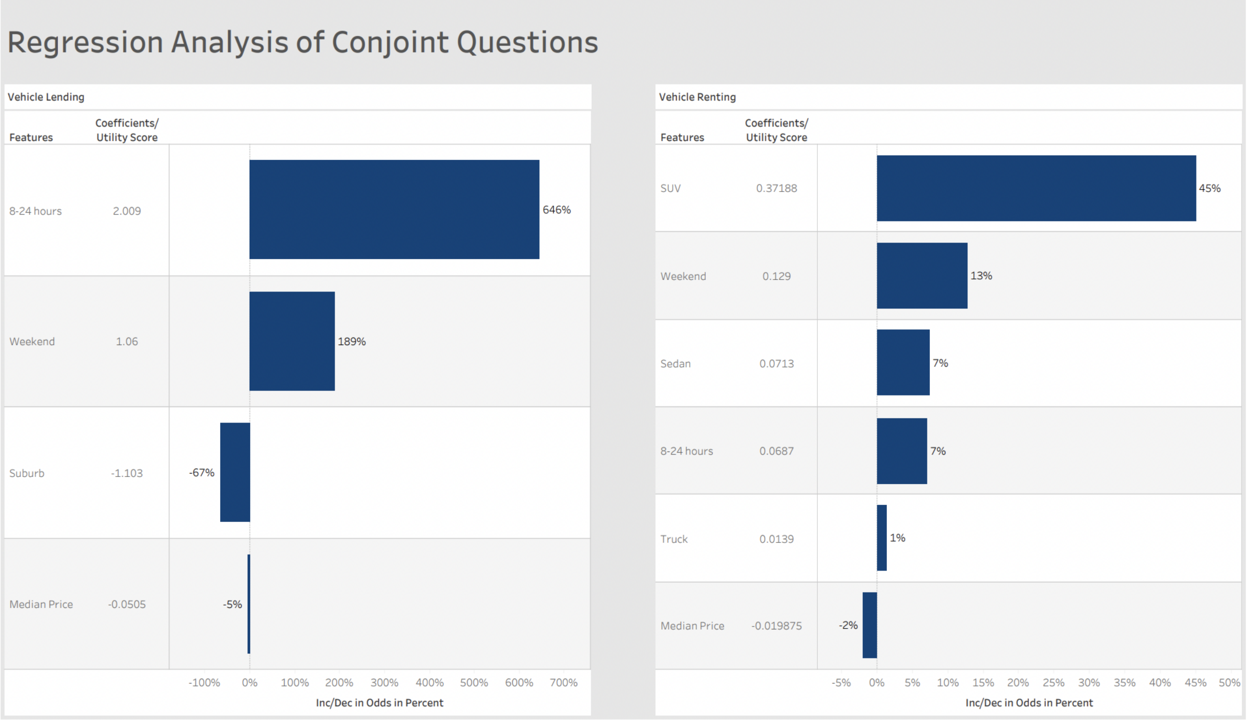
The last section of the survey was focused around conjoint style questions. We decided to incorporate this type of questioning in order to limit bias and to help us determine the consumer's sensitivity to different features. In these questions we gave both renters and lenders different scenarios to choose from. We gave them different vehicles, rental duration, time of the week and prices to choose from and they chose the combination they preferred. This helped us gain a better understanding of what was most important to the consumers and ultimately, how price sensitive they are. In order to determine this we conducted logistic regression on these results. This is discussed in more detail in the next section

Logistic Regression for Conjoint Analysis

The only available car type in the dataset was Acura RDX which is a type of SUV car. Therefore, we exclude this from the survey on the lending side. From the conjoint analysis, we calculated utility scores of each of the attributes using a logistic regression model. A utility score measures how much each attribute influences the customers' decision to select an alternative ("How to Interpret Partworth Utilities"). Although the survey data shows that 72% of people are willing to lend their cars through the platform, the detail is not sufficiently granular and does not give information about what attributes are important in making a lending decision. Instead, we can calculate the probability of participating in the platform using the total utility score.

The figure below shows the coefficients derived from the logistic regression model on the lender side dataset. We deliberately set the intercept to zero because, when all attributes are zeros, the utilities score should also be zero. These log-odds coefficients can be thought of as the utility score for each of the attributes. Notice that the 'features' column does not include city location or does not include weekday time. This is because, in creating dummy variables for categorical variables, one category must be dropped to prevent multicollinearity issues. Each of the dropped values has a coefficient of 0. (Kuila, n.d.). The numbers are relative to each other and can give insights about the ranking of feature importance. For example, duration within 8-24 hours increases the utility score by 2 on average. Customers value longer duration car lending more than shorter duration (2-8 hours). Similarly, if the lending location is in a suburb area as opposed to a city location, the utility score decreases by -1.1 utility score on average. Weekend lending time increases the utility score by 1.06. Another interesting insight is that lenders do not emphasize on the potential earnings. Another way to interpret these results is to take the exponential of the coefficients and interpret as an increase or decrease of odds of being selected for each of the features shown in figure X.

Similarly, the utility scores and the increase or decrease in odds percentage for the renters survey can be shown using the coefficients derived from the logistic regression model (Table X and Y).



Exponential of coefficients gives the increase or decrease in odds

As mentioned previously, the attributes measured for the renters differed slightly from the lenders-instead of location (suburb or city); we measured the type of vehicles to gain insights on the usage type. In table X, SUV is the most favorable type of car with a utility score of 0.37. This confirms that the RDX model in the Telematics dataset can be a good starting point for releasing peer-to-peer lending service. Like the lending side, the renting side also put emphasis on the duration and the time of the transaction.

Probability of lending calculation

We can use the following equation $P = exp(Total\ Utility) / exp(Total\ Utility) + 1$ to calculate the probability of lending.

As an example, we will select two random concepts to calculate the probabilities shown in the table below.

Concept	Suburb/City	Price	Weekend/Weekday	Duration	Total Utility Score	P(Lending)
1	Suburb	25	Weekday	8-24 hour	-0.38	41%
2	City	80	Weekend	2-8 hour	-3.02	4.7%

*Utility score for two random scenario*s

By changing the level in each attribute, the total utility score will be affected and thus increase/decrease the probability of lending. By using this method, we can see the maximum probability of lending by setting each level to determine the optimal total utility score.

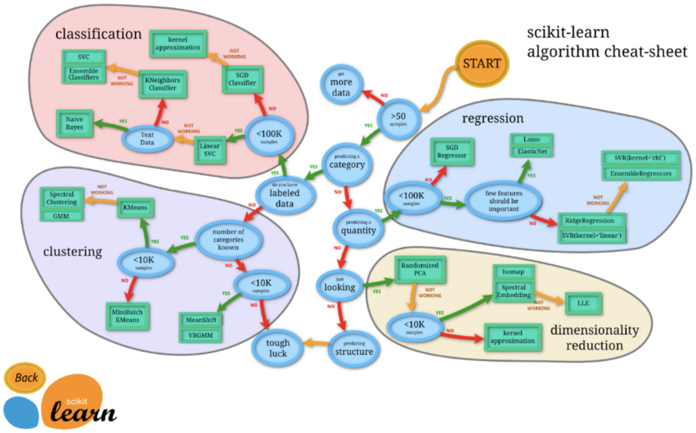
Suburb/City	Price	Weekend/Weekday	Duration	Total Utility Score	Max P(Lending)
City	20	Weekend	8-24 hour	2.04	88.49%

Maximum utility score scenario

Modeling Methodology

One of our main tasks was to identify which features in the Telematics dataset are most significant in determining a vehicle's dwell time. As a starting point, we decided to follow the dwell time model developed by the Berkeley team. However, rather than using only location data and a daylight variable, we included as many additional features from the dataset as we had post-data processing. Among these 45 selected features were the location clusters calculated above, time and duration variables, diagnostic warning variables, and several others. We performed multi-class classification, dividing the target variable (dwell time) into the three buckets defined in the existing dwell time model: 0–3 hours, 3–6 hours, and 6+hours.

Linear Models vs. Tree-Based Classifiers



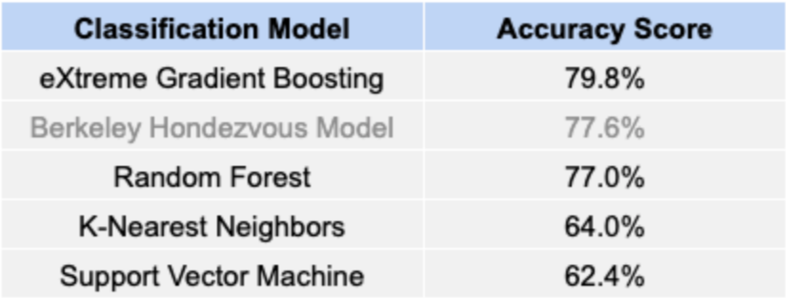
Scikit-learn Machine Learning Algorithm Flowchart
(Source: https://scikit-learn.org/stable/tutorial/machine_learning_map/)

In reference to the infographic, we focused our attention in the classification space and tested a variety of classifiers on the dataset. Linear models tend to be preferable when working with datasets containing fewer observations, whereas tree-based methods are highly scalable and work well with extremely large datasets. While the original dataset consisted of 25M rows, our resulting dataset after cleaning and processing included only 85K rows. As a result, we decided to evaluate both linear and tree-based methods in our analysis, in particular Support Vector Machines, K-Nearest Neighbors, Random Forest, and eXtreme Gradient Boosting.

Models like Support Vector Classifiers (SVC) and K-Nearest Neighbors (KNN) are linear separators in that they determine a boundary line or hyperplane to distinguish between classes. Because the problem we've defined involves more than 2 distinct classes, our methodology must allow us to map the data into a higher dimension. Although not all linear classifiers are capable of capturing non-linear relationships in this way, KNN does this natively while SVC achieves this by bringing in a kernel function, making them both preferable for our purposes. Furthermore, linear classifiers are often selected for their interpretability due to the availability of model coefficients.

Random Forests (RF) and eXtreme Gradient Boosters (XGB) are examples of tree-based models. These models are highly robust and can capture nonlinear relationships well. In addition to this, they are extremely effective in learning complex relationships that exist in high-dimensional data. The cost of this is that they are harder to interpret and can easily overfit to the data.

XGBoost Classification (Best Model)



Classification Model	Accuracy Score
eXtreme Gradient Boosting	79.8%
Berkeley Hondezvous Model	77.6%
Random Forest	77.0%
K-Nearest Neighbors	64.0%
Support Vector Machine	62.4%

Comparison of Model Results

In particular, we leveraged an eXtreme Gradient Boosting model, a decision tree classifier that employs an iterative process used to minimize loss, called gradient descent. This model attempts to learn as it goes, correcting past mistakes to improve its performance. When used for classification tasks, XGBoosting methods are often successful in achieving high accuracy and low bias with low computation time. We were able to validate this hypothesis with a test accuracy of 79.8%, which is a 2.2% improvement over the previous 3-class dwell time classification model.

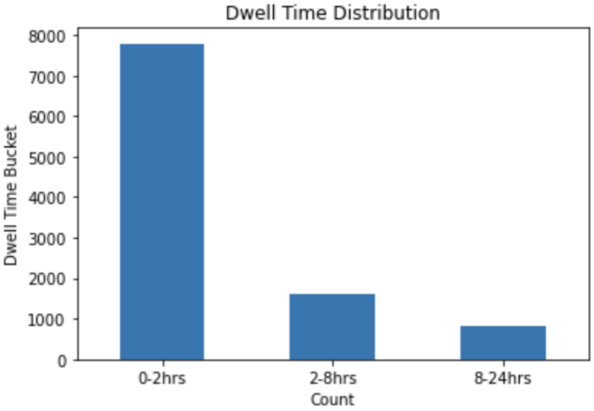
Drawbacks of Chosen Dwell Time Buckets

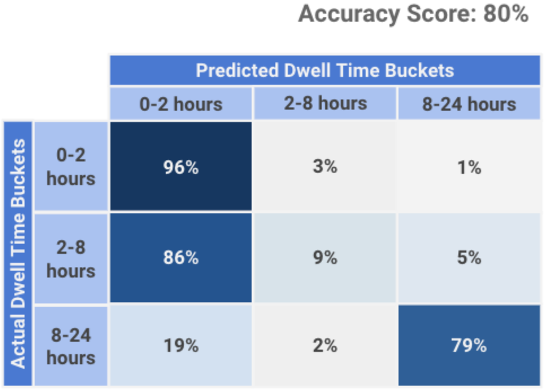
Our best-performing XGB model provided great results in terms of overall test accuracy. However, digging deeper into the results quickly revealed that this model failed to be ideal. The distribution of our target variable shows that a large majority of the observations had short dwell times, between 0 and 3 hours, whereas vehicles that dwelled for 3–6 hours comprised a meager fraction of the dataset. As expected with class imbalance, the confusion matrix for this XGB model shows that the high accuracy achieved is mainly attributed to the fact that our model correctly predicted observations in the 0–3 hour class.

Addressing the Drawbacks

We decided to select new dwell time buckets that better apply to our two business cases: 0-2 hours, 2-8 hours, and 8-24 hours. Instead of having the top-level bucket be unbounded, we limited the dwell times to 24 hours because our P2P business case focuses on shorter-term rentals.

While the new dwell time buckets were more intuitive to our use cases, the data continued to suffer from severe class imbalance. We quickly discovered that the accuracy score wasn’t a reliable metric to observe. Instead, we want to focus on the recall score, which measures the percent of correctly-identified positive cases. This is because we want to ensure that the model predicts the correct dwell times for each trip.

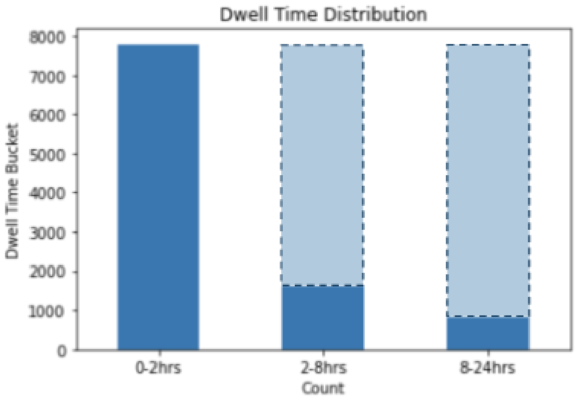


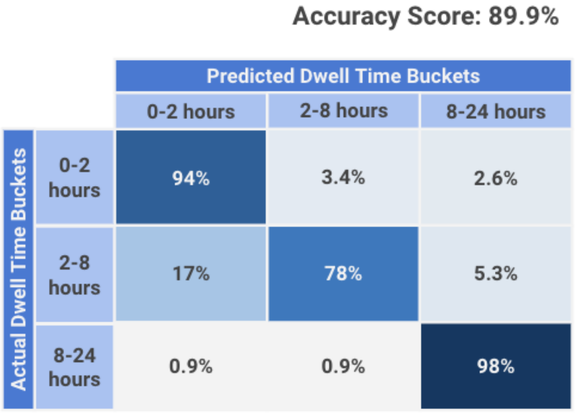


Accuracy Score: 80%

		Predicted Dwell Time Buckets		
		0-2 hours	2-8 hours	8-24 hours
Actual Dwell Time Buckets	0-2 hours	96%	3%	1%
	2-8 hours	86%	9%	5%
	8-24 hours	19%	2%	79%

SMOTE (<u>S</u>ynthetic <u>M</u>inority <u>O</u>versampling <u>TE</u>chnique)
We employed an oversampling technique called SMOTE to address the issue of class imbalance. This technique synthesizes new observations for the minority classes using k-Nearest Neighbors methodology. By doing this, we were able to increase the number of observations in the 2-8 and 8-24 hour buckets and even out the dataset. Putting this through the XGBoost model returned much more optimal and better balanced recall scores.





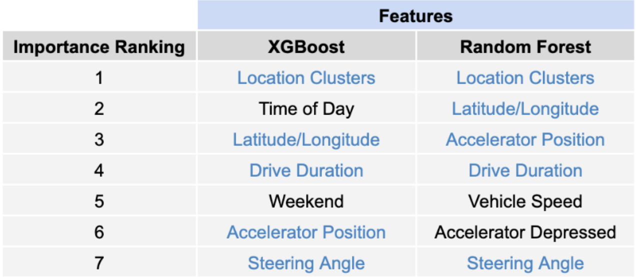
Accuracy Score: 89.9%

		Predicted Dwell Time Buckets		
		0-2 hours	2-8 hours	8-24 hours
Actual Dwell Time Buckets	0-2 hours	94%	3.4%	2.6%
	2-8 hours	17%	78%	5.3%
	8-24 hours	0.9%	0.9%	98%

Feature Importance via SHAP

While our tree-based models outperformed the linear models significantly, these methods suffer from being black boxes, making them difficult to explain and interpret. In order to circumvent this drawback, we leveraged SHapley Additive exPlanations, otherwise known as SHAP, to determine feature importance.

The concept of SHAP is based upon game theory and machine learning, using the outcome and features of the model as its 'game' and 'players' to "quantify the contribution that each feature brings to the prediction made



	Features	
Importance Ranking	XGBoost	Random Forest
1	Location Clusters	Location Clusters
2	Time of Day	Latitude/Longitude
3	Latitude/Longitude	Accelerator Position
4	Drive Duration	Drive Duration
5	Weekend	Vehicle Speed
6	Accelerator Position	Accelerator Depressed
7	Steering Angle	Steering Angle

Feature Importance Rankings (variables that are mutually important are highlighted in blue) by the model”. This methodology focuses on local interpretability, meaning that it observes the model outcomes at a per-observation level. Aggregating these values allows us to evaluate feature significance at various levels of granularity.

At the full-dataset level, the Shapley values suggest that there are a number of mutually significant features between the XGBoost and Random Forest models. In particular, location variables such as the DBScan-predicted clusters and latitude/longitude have the greatest impact on the dwell time.

Cost-Benefit Analysis

Vehicle Maintenance Service

In this section, the cost benefit analysis for car maintenance service is presented. The approach of the calculation will be slightly different from what is reported in the peer-to-peer lending service. However, the idea is similar – we want to show that using dwell time prediction as mentioned in the business case would provide some positive value and is worthwhile continuing as a more in-depth analysis.

In order to calculate the cost-benefit, we make a few assumptions which are based on the survey result:

1. The subscription fee will cover labor, service and material cost and will be $200 per year on average.
2. Service subscribers will get two routine maintenance services, meaning customers will pay about $100 each time they accept the service.
3. Since our business model requires potential customers to either accept or decline the service, we leverage Honda app to notify the customer about the service. We assumed that the marketing cost will be $0.5 per push notification.
4. From the survey result, we know that about 76% of the customers are interested in using the service, so this will be our probability of accepting the service.
5. Customers will receive a cash discount/gift card of about $20 per hour if the service exceeds the agreed-upon duration. We will assume that overtime is less than or equal to 1 hour.

	Predicted suitable	**Predicted not suitable**
Actual suitable	**Action** Honda notifies the customer. Customer has 76% chance of accepting the service. **Expected Cost** $0.5 marketing cost **Expected Benefit** $0.76 * ½*subscribing fee earning **Expected Profit** $75.5	**Action** Honda does not notify the customer. Customer has 76% chance of accepting the service. **Expected Cost** $0.5 marketing cost $0.76 * ½*service fee earning **Expected Benefit** $0 **Expected Profit** -$75.5
Actual not suitable	**Action** Honda notifies the customer. Customer has 76% chance of accepting the service **Expected Cost** $0.5 marketing cost $20*Overtime(Hour) penalty Cost **Expected Benefit** $0.76 * ½*subscribing fee earning **Expected Profit** $76 - 20*(1) - 0.5 = $55.5	**Action** Honda does not notify the customer. **Expected Cost** 0 cost **Expected Benefit** 0 earning

Cost benefit analysis framework

	Predicted suitable	Predicted not suitable
Actual suitable	**16001**	3215
Actual not suitable	1041	**8516**

Confusion Matrix from the ML model

	Predicted suitable	**Predicted not suitable**	**Total Earned**
Actual suitable	**$1,208,075**	**-$242,733**	**$965,342**
Actual not suitable	**$57776**	**0**	**$57,776**
		Total Expected Profit	**$1,023,118**

The total potential expected profit created for the car maintenance service using the ML model is $1,023,118.

Peer-to-Peer Lending Service

We can now conduct a cost-benefit analysis from our machine learning model for the peer-to-peer lending case. The assumptions we made in order to make this calculation are:

1. We will assume a 40% takeaway fee for Honda, meaning the customer will earn 60% from using the platform.
2. The cars for lending and renting are the RDX model.
3. Only the base price is taken into consideration. In actual business settings, there will be additional fees incurred from using the platform.
4. For 2-8 hours, we will assume a base price of $20.
5. For 8-24 hours, we will assume a base price of $50.
6. At the start, we will only offer a service in the City area.
7. Assume that profit is realized when lenders decide to put their cars on the platform, i.e.there will always be renters who will use the service
8. Customers will not accept if their actual dwell time is within 0-2 hours.
9. The probabilities derived from the logistic regression on both lending and renting side are representative of the probability of someone lending their cars as well as the probability of someone renting the available cars on the peer-to-peer car platform.

The table below shows two product concepts that will be used to illustrate the costs and benefits of the machine learning model. From the survey result, City Location and Weekend were the most preferred choices for both lending and renting sides, thus we selected both concepts to have these levels. One of the concepts will have a short term (2-8 hrs) duration of lending/renting and the other will have a long term duration (8-24 hrs). With these attributes selected, the total utility scores were calculated using the coefficient from the logistic regression. We then converted the scores into the probability of lending and renting.

Suburb/City	Price	Car Type	Weekend/Weekday	Duration	Total Utility Score	P(Renting)	P(Lending)
City	50	SUV	Weekend	8-24 hour	0.51	47.5%	62.0%
City	30	SUV	Weekend	2-8 hour	-0.47	39.5%	38.5%

Probability of lending calculation for two different scenarios (used for illustrating the cost-benefit)

With the probability of lending *P(lending)* calculated, we constructed a cost-benefit table below comprising the potential actions in each of the scenarios. In this table, we only consider the cost-benefit for the weekend scenario because of customer preferences taken from the utility scores. For 0-2 hours, Honda will not notify the customers to place their car on the platform. If correctly predicted, the expected earning would be $0. However, if incorrectly predicted, there will be opportunity costs incurred to Honda because customers are inclined to place their cars on the car sharing platform. For the predicted value that falls under the 2-8 and 8-24 hours bucket, Honda will make an offer to the customer. Customers will only be likely to place their car on the platform as long as it is within their dwell time. Therefore, the potential earnings can be calculated as the multiplication of lending price, and the probability of lending P(lending | hr = 2-8) or P(lending | hr = 8 - 24) minus the cost for notifying the customers. For incorrect predictions, we assumed that the only cost incurred to Honda is the marketing cost. This includes the app maintenance fee shared across all users. We assumed that this theoretical marketing cost is $0.5 per push notification.

Predicted/Actual Dwell Time Buckets	**0-2**	**2-8**	**8-24**
0-2	**Action** Honda does not offer the service to customers **Expected Cost** $0 **Expected Benefit** $0 **Expected Profit** $0	**Action** Honda ask customers to place their cars on the car lending platform **Expected Cost** $0.5 marketing cost **Expected Benefit** $0 **Expected Profit** $-0.5	**Action** Honda ask customers to place their cars on the car lending platform **Expected Cost** $0.5 marketing cost **Expected Benefit** $0 **Expected Profit** $-0.5
2-8	**Action** Honda does not offer the service to customers **Expected Cost** $0.5 marketing cost **Expected Benefit** -0.385 * 0.475 * $30 = -$5.5 **Expected Profit** -$6	**Action** Honda ask customers to place their cars on the car lending platform **Expected Cost** $0.5 marketing cost **Expected Benefit** 0.385 * 0.475 * $30 = $5.5 **Expected Profit** +$5	**Action** Honda ask customers to place their cars on the car lending platform **Expected Cost** $0.5 marketing cost **Expected Benefit** 0.385 * 0.475 * $30 = $5.5 **Expected Profit** +$5
8-24	**Action** Honda does not offer the service to customers **Expected Cost** $0.5 marketing cost **Expected Benefit** -0.62* 0.395 * $50 = $12.25 **Expected Profit** -$12.75	**Action** Honda ask customers to place their cars on the car lending platform, the service to customers **Expected Cost** $0.5 marketing cost **Expected Benefit** 0.62* 0.395 * $50 = $12.25 **Expected Profit** +$11.75	**Action** Honda ask customers to place their cars on the car lending platform **Expected Cost** $0.5 marketing cost **Expected Benefit** 0.62 * 0.395 * $50 = $12.25 **Expected Profit** +$11.75

Cost-benefit analysis for the weekend model

Confusion matrix (only weekend and capped at 24 hours)

	0-2 hrs	2-8 hrs	8-24 hrs
0-2 hrs	1451	52	40
2-8 hrs	267	1247	63
8+ hrs	14	13	1508

Confusion matrix for weekend model (actual time buckets in rows)

Expected profit for the Weekend Model
Using the cost-benefit table and the confusion matrix, we perform element-wise matrix multiplication to get the total expected profit.

Predicted/Actual Dwell Time Buckets	**0-2**	**2-8**	**8-24**	**Total Expected Profit**
0-2	$0	**-$26**	**-$20**	**-$46**
2-8	**-$1331.33**	**+$6217.85**	**+$314.13**	**+$5200.66**
8-24	**-$164.43**	**+$152.69**	**+$17711.46**	**+$17699.72**

The total potential expected profit created from the weekend model is 2. With 40% takeaway, Honda can earn up to $9141.75 and the value created for customers is $13712.62.

Conclusion

The two proposed business models integrated with the predictive models show the opportunities for 99PLabs. With the model and survey results and the research conducted in this project, both business models present promising opportunities. The modeling and survey validated the market for 99PLab. A risk framework consisting of FMEA tables and cost benefit analysis were developed for quantifying the model risks as well as possible methods to mitigate them.

There is, however, significant more work that needs to go into each of the business cases. The 2 weeks data were collected in 2018, and the distribution of the data may have changed. Another challenge is that with constant incoming data from sensors, a strong data engineering infrastructure will need to be able to handle stream data before 99PLabs can consider deploying the model into production.

References

"How to Interpret Partworth Utilities." n.d. Conjoint.ly. Accessed May 1, 2022. https://conjointly.com/guides/how-to-interpret-partworth-utilities/.

Kuila, Ritvik. n.d. "Modeling Consumer Decisions: Conjoint Analysis | by Ritvik Kuila." Towards Data Science. Accessed May 1, 2022. https://towardsdatascience.com/modeling-consumer-decisions-conjoint-analysis-f4eda531ecf6.

Mazzanti, Samuele. 2020. "SHAP Values Explained Exactly How You Wished Someone Explained to You." Towards Data Science. https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30.