Identifying tourists profiles using data from Location-Based and Travel Social Networks

Lucas Félix
Univesidade Federal de Minas Gerais
Brasil
lucas.felix@dcc.ufmg.br

Jussara Almeida
Univesidade Federal de Minas Gerais
Brasil
jussara@dcc.ufmg.br

KEYWORDS

Tourism, Recommendation, Data Mining, Machine Learning

HIGHLIGHTS - WORK REQUIREMENTS

- Presentation Link
- Alternative Link - YouTube Option
- We propose two different strategies to perform trip purpose classification, based on text classification and ML classic approaches.
- We perform system comparison by statically comparing the differences between the approaches, we use a 5-fold, with 95% confidence t-paired test, and Bonferroni correction.
- We perform a factorial design with a Random Forest model, identifying the features that mainly impact the prediction of trip purpose
- We perform a deep analysis of the data to verify if the Logistic Regression method is suited for our problem
- We perform a deep characterization of tourists by making a clustering analysis, and identifying the difference in these users' behaviors.

1 INTRODUCTION

Before the internet, the spread of tourism information relied on two main ways: personal experiences exchange, from one person to another, and paper-based recommendations, made by guides and magazines [24, 26]. In both cases, these recommendations influenced other people's choices [26, 45]. Nevertheless, these were generally generic suggestions based on a single person's opinion. In a scenario as intangible and heterogeneous as tourism, many voices can bring many points of view [3]. However, these recommendations were restricted to the personal social circle.

In this sense, the internet was one of the catalysts for tourism development [45]. The advance in connectivity directly impacts this industry, which has increased its value since the 2009 global crisis [22] until the 2020 COVID-19 pandemic, showing how this was one of the services that most benefited from the internet expansion [45].

Through the internet, users can now gather in forums, communities, travel blogs, and *Social Networks* (**SNs**). These platforms made it possible to seek *Points of Interest* (**POI**), provide and receive information about places, sharing experiences from a single point of view. Studies show that 80% of American travelers use SNs while traveling, and more than half of this percentage share their journey information with their contacts [45]. These studies also show how SNs are most the adopted between travelers[45].

Between the platforms available on the web, three different types of SNs are generally used: i. *General-Purpose SN*; ii. *Location-Based SN* (**LBSN**); and iii. *Travel SN* (**TSN**). General-purpose SNs, such as Twitter, Facebook and Instagram, usually contain general information and not only tourism-related content. Although these platforms enable their users to geo-localize the shared content, these are usually restricted to the user's social circle and are not structured to easily identify the tourist opinion about the place through a rating or a review.

The two other types of platforms – LBSN and TSN – are more focused on tourism and have been extensively used in the literature in tourism works [22, 24]. With different purposes, LBSNs (e.g., *Yelp* and *Foursquare*) are more centered on the user's current location sharing their footsteps. In contrast, the TSNs (e.g., *TripAdvisor*) are responsible for joining in one platform several pieces of information about locations, accommodations, transport, food, attractions, and services [45], enabling their users to better describe the trip made. Furthermore, TSNs also offer the user the possibility to plan trips and get recommendations about users with the same taste [23] while comparing prices and looking at different perspectives on a place.

Even though TSN and LBSN platforms concentrate most of the information needed to plan a trip to one place, users still experience difficulty with the amount of available data, precluding to distinguish which option (or set of options) is the best [34]. Therefore, the *Recommender Systems* (**RS**) are usually applied to tackle the information overload problem, focusing on suggesting POIs based on the user history. RS approaches reduce the number of options for the users and enable them to choose between a smaller (and possibly better) set.

Different from conventional recommendation scenarios (e.g. movie recommendation), POI recommendation is considered to be a harder task, due to geographical (e.g. physical constraints between the POIs) [7], social (e.g. friends can influence on the user visit) [4, 50] and temporal (e.g. places are more searched on summer) influences [31]. Besides, the number of instances tends to be much smaller, sparser, and noisier than in book and movie scenarios [5, 44]. Also, the heterogeneity of information in social networks describes the user activity from a variety of perspectives and through different types of data (e.g. photos, text, check-ins), which lead to a vast literature that focuses on different approaches for accomplishing the POI recommendation task [8, 14, 49].

Due to the many aspects that make the POI recommendation scenario harder, works on the literature use contextual information to leverage recommendations quality. As stated in [40], additional data allows for more accurate recommendations than traditional methods, thus, works exploit geo-coordenates [8, 29], social ties [43, 49], places category [28, 48] and temporal information [49] to enhance predictions quality.

One additional piece of information that has been little exploited in tourism recommendations is the tourist profile. Besides being a difficult task to properly characterize users, the few works that do so, make assumptions that do not correspond to tourists' real behavior, hence, biasing their conclusions. Illustrating that, authors in [13], proposes a tourists profile characterization using LBSN data. In this work, authors remove from their dataset travels with less than 7 days, aiming to remove for their evaluations work-related travels. In this scenario, authors justified that work-related travels are not interesting, due to the fact that they are focused on tourists. Nevertheless, in [15], authors states that most trips range from 3-8 days, which makes the characterization made in [13] leave aside most of the trips made by users, hence having little practical value. While in [2], authors only focuses on restaurants visitations profile. Besides that, to the best of our knowledge, only two work on the literature explore tourists profiles using real data and proposing a taxonomy for the users [2, 13].

Thus, in this work, we propose a methodology to characterize tourist profiles, aiming to fill some gaps left in the literature. To guide our work, we aim to answer the following **Research Questions (RQ)**:

1. Is it possible to properly identify which type of trip a traveler is making (work or leisure)?
2. What are the main differences between the users profiles identified by our methodology?
3. Does the taxonomy proposed in [2, 13], still holds up when using a properly pre-processed dataset?

To answer the above RQ, we propose a methodology that first address the issue of identifying the trip type, with two different strategies: A text classification method and classic ML model (struct data, e.g. Random Forest). Then, we use this model to classify the travels made by users of LBSN, in special, Yelp. We use an unsupervised machine-learning model to identify clusters of tourists, characterizing and classifying these profiles. We believe that with the proposed methodology, RS algorithms can use travelers' profile information to leverage their recommendations.

2 RELATED WORKS

The goal of this research is to characterize the tourists profile and increase the quality of RS suggestions. This section discusses the literature on datasets used on tourism RS, a brief overview on text classification, clustering analysis, and previous attempts to characterize travelers.

2.1 Datasets

The act of traveling is part of human history and has become much easier in the last years due to the increase in mobility and information access [20]. With the web popularization, people were encouraged to search about new destinations that they could visit. Besides, users were now able to share their experiences with other people through social networks and traveling specific platforms [45]. The popularization of these platforms made studying tourism a more straightforward task given the amount of data available [3, 26].

Currently, there are available in the literature datasets extracted from *Location-Based Social Networks* (**LBSNs**), such as *Gowalla*, *Foursquare*, and *Yelp*, and *Travel Social Networks* (**TSNs**), such as *TripAdvisor*, *Real Travel*, and *Travello*, which are mainly used for the task recommendation and route generation.

LBSNs are social platforms where users share in real-time their geospatial location and visit timestamp, composing a visit check-in [4]. Users of these platforms show a pattern of use on a regular daily basis, and about 90% of their transitions are between places are within the 50 km range [31]. Corroborating with that, the work of [19] shows that most of the locations are followed by fewer than 5 different locations consecutively, showing that users have a similar visitation pattern.

Considering the LBSNs available datasets, two different features can be found. First, datasets from *Foursquare* and *Gowalla* present a small set of features, composed of the User identifier, POI identifier, POI category, and timestamp. In this case, works that employ Foursquare and *Gowalla* data assume that if a user visited a place, then the user has liked it since no explicit feedback is available [24, 46, 47]. On the other hand, *Yelp* data are more suited for the recommendation, given that additionally to check-in data, it enables the users to share a detailed evaluation for each POI (content information).

Besides LBSNs, TSN data is also widely used in tourism evaluation [22, 42]. TSNs are social platforms focused on tourism, where users can share visited places and reviews with other travelers. Unlike LBSNs, TSNs focus on a "couch review", i.e., the travelers make their review after visiting a place, in many cases months after that visit. In this scenario, the check-in time is unavailable, only the month and year of visit. However, the information presented in a TSN review is fine-grained. The user supplies explicit ratings to aspects such as the trip type (e.g. work-related, leisure-related), place service, cleanness, and location, providing a written review. The main advantage of TSN datasets over Yelp data is that place owners also share their profile, defining features such as price and opening hours and having a closer relationship with travelers.

In this work, we use both TripAdvisor and Yelp data. TripAdvisors' data is used to extract the users' trip type, then we use transfer learning to identify Yelps travels type through users' review. Over Yelp data we perform the users' characterization, making a smart pre-processing, defining which trips are work-related and which are not.

2.2 Text Classification

Automatic Text classification (**ATC**) consists on the task of automatically detecting the document's class. The classes in each dataset can vary according to the scenario, allowing this type of technique to be applied in different scenarios as relevance feedback, sentiment analysis and product reviews [10].

ATC has recently experienced advances due to Transformer-based deep learning approaches (neural approaches), being currently the state-of-the-art on this type of task [11]. Models in this class are usually divided in two main steps: (i) pre-training: where the model learns the weights using an unsupervised task, and (ii) fine-tuning: where a supervised task, learns the specifics of the labeled datasets. This step allows further adjustments on the neural network weights, enabling the model to be domain specific. The success of these techniques is justified [32] by two main reasons:

first, the amount of data used to pre-train these models, and the second is the possibility of adapting the pre-trained model for specific domains, by simple adjusting the tuning of the last layers of the neural network.

Considering the tourism scenario, few works have applied neural approaches on this context. In [10], authors use a Yelp dataset in a comparative study aiming to evaluate text classification techniques, in special neural approaches. In [25], authors uses summarized review data to classify POIs' category by using a Long-Short Term Memory (LSTM) technique. Lastly, in [16], the authors applied aspect based sentiment analysis to identify words that are associated with a positive/negative review.

In this work, we use ATC to identify a user trip type from their review. To the best of our knowledge, this is the first work to apply a text classification model in such a context. We believe that this model enables researchers to better identify patterns in travels, and consequently travelers.

2.3 Clustering Analysis

Clustering is a task of unsupervised machine learning, that enables to find data partitions, given a value $k \in \mathbb{Z}^+$, where k is the amount of data partitions wanted. In this scenario, the data within each group more similar than the data than outside the group [36]. The main advantage of this type of technique is that it is context-agnostic, allowing it to be applied in several contexts, as in social [17], environment [18], and scholar analysis [33].

Between the several techniques available in the literature, k-Means, is the most popular approach [1], due to it simplicity, interpretability, and the fact that it can be easily modified to meet specific constraints and contexts. However, one of the difficulties in this type of technique is to define the value of k, due that this parameter directly impacts on how cohesive the points in a cluster are. Thus, to identify the clusters' quality, one technique used is the silhouette index, which works by exploring the mean distance between the data within a group, and the closest data outside the group. This approach is often used by its quality in different scenarios [38].

Considering the tourism context, several papers made use of clustering analysis in their evaluations. First, in work similar to ours [13], the authors use k-Means to identify different travelers profile within the Foursquare dataset. In [39], the authors use k-Means to identify if a visitor is a city residential, or a visitor.

In this work, we use clustering analyzes to identify profiles of travelers in LBSN datasets, similar to [13]. Different from then, we use a different set of features to characterize the travelers, also using a less aggressive pre-processing, aiming to identify hidden patterns that were not observed. We believe that with such characterization, we can enhance RS suggestions.

2.4 Tourists Characterization

Tourism characterization as been widely studied in the literature. From psychology [6] to computer science [2], several methodologies were applied to identify travelers classes. Nevertheless, most of these works are theoretical models [6, 9, 35], and were not observed with real data. With the use of SN by tourists, a massive amount of content were created, making possible for tourists characterization studies as [2, 13]. While the work of [13] makes assumptions in their pre-processing that do not hold on the real-world, the work of [2] does not evaluate human-mobility aspects of the travels, due to the fact that TripAdvisor data is used. Complementary, the authors in [2], only focuses on restaurants' evaluation, while many different categories of places are available for a user to visit. With a different focus, in [37], authors proposes a methodology for comparing cities by their users pattern of visits. The proposed technique, called city image, can also be applied in different contexts, for example, to compare different users types, showing the differences in their visitation patterns.

To the best of our knowledge, these are the only two works that defines a taxonomy to differentiate users profiles using SN data to validate [2, 13]. However, has pointed before, these works lack deeper evaluation of the users to properly characterize them. Thus, in this work, we propose a profound analyzes of tourists, using different techniques to properly characterize and differentiate users profiles, defining a taxonomy for users with different behaviors.

3 PROPOSED STRATEGY

In this section, we present our proposed methodology aiming to characterize the tourists profile and increase the quality of RS suggestions. As our proposal is dependent on specific data, we first introduce the data collections used on our analyses, then, we introduce the two major steps that compose our methodology: (*i*) trip purpose classification, and (*ii*) clustering analysis.

3.1 Data Collection

Most works in POI recommendation use LBSN data in their evaluations, given the availability of these datasets in the literature [8, 40, 43, 49]. However, as previously discussed, most LBSN datasets do not present explicit feedback and especially with quality. From the datasets available in the literature, only a few present essential features for real-world evaluation, such as the POI working hours, availability, and cost. In this case, data from TSN, like TripAdvisor, are more suited for such analyzes. Besides containing more information, TripAdvisor is currently the most popular travel website [22], with about 390 million monthly unique visitors. Similar to TripAdvisor, Yelp, as mentioned in Section 2, is one of the few LBSN datasets that contain fine-grained data. However, it lacks features that help to characterize the user's trip, as the trip type/trip purpose. With such information, it would be possible to avoid assumptions that mischaracterize users' behavior. Thus, to perform a proper tourist profile, in this work, we use both TripAdvisor and Yelp data. We use TripAdvisor review data to train a classifier capable of distinguishing work and leisure-related travels. Then, we use the trained model to discriminate travels in Yelp, hence producing a proper traveler characterization.

It is noteworthy, that unlike the work [13], that performs such characterization in a Foursquare dataset, we use Yelp data, due to the presence of quality explicit review on the data.

To extract TripAdvisors' data, we developed a crawler responsible for automatically browsing the website collecting all users' available content and POIs. The collected data was firstly in an unstructured format (e.g., HTML). Then, a parser was developed to retrieve the content within each page, pre-process, and store it in a semi-structured format (e.g., CSV) data.

Our collection was initially focused on users from five different touristic cities worldwide and their complete visit history (including other cities): Tiradentes and Ouro Preto (Brazil), San Gimignano, Cannes, and Ibiza [1]. Since our dataset possessed users' complete history, containing the visits to numerous different cities, in this work, we focused on the English reviews available [2] that are associated with the trip label (e.g. family, romantic, friends, work-related, alone) [3]. We focused the analysis on English reviews due to the fact that the text classification algorithms employed in our methodology demonstrate better performance with English rather than in other languages [41]. To identify the English reviews on the datasets used in this work, we adopted the pre-trained model proposed in [21].

To augment our dataset, we labeled visits occurring within the same city and time period (month and year) that had a missing label. To assign a label, we used the available classification from another POI visited by the same user in the same city and time period.

Considering our aim is to define leisure and work-related travels to characterize travelers further, initially, we modeled our problem as a binary classification problem by aggregating non-work classes as leisure, hence leaving only two classes. Following the aggregation process, it was observed that 87.67% of the instances consisted of leisure-related reviews, whereas 12.33% were categorized as work-related reviews. This indicates an imbalanced dataset context, where most instances pertain to leisure-related reviews.

Table 1 presents a summary of TripAdvisor data used in this work.

Dataset	**# Instances**
TripAdvisor Complete	11, 443, 663
TripAdvisor English	2, 434, 252
TripAdvisor English w/ classes	639, 997

Table 1: TripAdvisors' data summary

Yelp made available in 2016 their dataset, allowing it use for personal, educational, and academic purposes [4]. As stated, Yelp dataset has the advantage of having written reviews, and the users' check-in data, allowing to properly identify the users' opinions and mobility patterns. As the ATC model used in this work is trained in English, we also filtered the Yelp datasets by analyzing only English reviews. In Table 2, we present the summary of Yelp data used in this work, it is possible to notice that only a few of the instances are not in English, this is due to the fact that the Yelp dataset is mostly located in North-America.

Dataset	**# Instances**
Yelps Complete	8, 635, 403
Yelps English	8, 625, 211

Table 2: Yelps' data summary

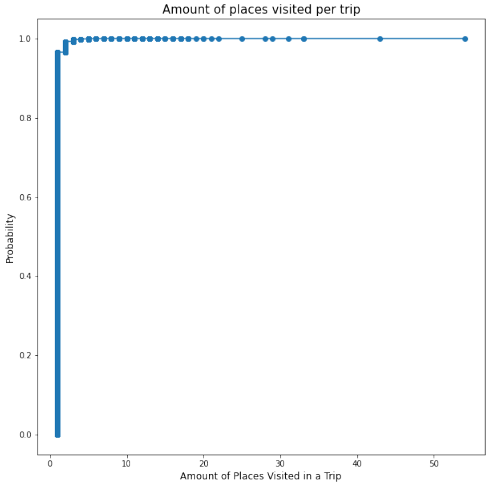

Figure 1: Amount of places per trip

After filtering the English reviews in the Yelp dataset, we group the travels made by the users, aiming to identify continuous check-ins that belonged to the same trip. To define the trips, first, we defined a users' home city, as the city where the users made the majority of their check-ins. Then, we ordered the users' check-ins considering the following grouping criteria: While the user is making continuous check-ins in places out of town and the maximum time distance between these check-ins is about 24 hours, then all of these check-ins are grouped as a trip. However, if a user makes a check-in in his hometown, or the time between the check-ins out-of-town is over the 24-hour range then, it is considered a new trip. Figure 1, presents the cumulative percentage of places visited per trip. It is possible to notice that most of the trips only have one visited place. As discussed in Section 2, this is due to the fact Yelp users use the platform to make check-in not only in touristic POIs but also to evaluate day-to-day places, such as mechanical workshops, and car dealerships. In the end, where identified near 1.8 million travels made by users.

3.2 Work or Leisure Classification

For the step of work or leisure classification, we test two different approaches to properly classify a trip purpose. First, we evaluate a text classification approach training the model with TripAdvisor-labeled data, and then labeling Yelp data. Our second approach is based on Machine Learning (ML) general tabular structured models such as Logistic Regression, and Random Forest.

Text Classification. In the ATC block of our modeling, we aim to classify user reviews, identifying whether their travels were work or leisure related. To accomplish so, we compare three different neural approaches that are state-of-the-art on the text classification task, achieving the best results in the benchmarks used by researchers [10], namely we use Bert [12], RoBERTa [30] and BART [27].

As mentioned, the above tasks are Transform-based deep learning techniques, been the main two steps of these algorithms are described in Section 2. RoBERTa and BART are enhancements to the

[1] TripAdvisor Complete Dataset
[2] TripAdvisor English
[3] TripAdvisor English w/ classes
[4] https://www.yelp.com/dataset

procedure defined by Bert. While RoBERTA, focuses on properly fine tune the parameters on BERT, especially the weights on the last layer of the neural network[30], BART works by introducing noise on the original text and trying to reconstruct the original text [27].

To compare the effectiveness of each approach, we follow the procedure defined in [10]. We evaluate the effectiveness of the proposals using Macro and Micro Averaged F1. The experiments were executed using a 5 fold cross-validation procedure. To compare the cross-validation results, we evaluate the statistical significance using a paired t-test with 95% confidence. Even though we have few algorithms to test, we apply the Bonferroni correction to account for the multiple tests.

As mentioned earlier, our dataset exhibits a significant skew, making it challenging to obtain accurate predictions for both classes. To address this issue, we implemented a class balancing procedure during the training phase for each fold, in which we randomly chose N_{mino} instances from the majority class, ensuring equilibrium with the N_{mino} instances derived from the minority class. During the testing phase, on the other hand, we maintained the overall distribution of the datasets to ensure a realistic representation of the actual data.

As previously stated, our aim with the ATC block is to apply the model trained with TripAdvisor dataset, to Yelp dataset, hence, identifying work and leisure travels. As the Yelp dataset does not have a label, it is not possible to quantify the quality of the transfer learning. Thus, to quantify the effectiveness of our proposal, we manually classified 1000 random select Yelp instances. We use this manual-labeled dataset as a ground truth, a compare the text classification proposes in terms of Micro-F1 and Macro-F1.

***ML-based trip classification*.** Complementary, we extracted features of the travels made by users and, evaluate different ML classification models to identify the main factors that impact a work and leisure trip. For this step, we train the model directly on Yelp dataset, due to the fact that TripAdvisors' data do not present the features set needed to posteriorly perform such classification on Yelp. To evaluate the model, we used a procedure similar to the one used on the ATC block of our methodology. Below, we list the features used in the classifiers:

- **Visit Month:** Visitation trip month. (One-hot-encode)
- **Amount of Places:** The number of places visited in a trip.
- **Travel Duration**: Trips' length in days
- **Hotel Distance:** Mean distance between the places visited and the hotel that the user is staying in.
- **Hometown Distance:** Mean distance between the places visited and the users' hometown.
- **Radius of Gyration:** Mean distance between all places visited.
- **Percentage of Business Hours**: Percentage of check-ins made during business hours.
- **Check-in frequency:** The number of check-ins made during the trip divided by the travel duration.
- **Check-in Duration:** Mean time between two consecutive check-ins.
- **Check-in Time:** Mean hour that check-in is made by the user.
- **Stars:** Mean amount of stars given to each place visited.
- **Places visited category:** Amount of places per category visited (e.g. Restaurant, Hotel, Attractions, and Automotive)

For this task, we used three different strategies: Logistic Regression, Random Forest, and a Dummy Classifier.

3.3 Clustering Analysis

In the Cluster Analysis step of our methodology, we aim to identify clusters of travelers with similar behavior, characterizing the users, and then verify the taxonomy proposed in [13]. In this work, we use k-means to characterize the Yelp dataset. As stated in 2, we opted for k-means due to the techniques' simplicity, and good quality results in different scenarios. To define the clusters' quality, we use the silhouette method.

4 EXPERIMENTAL EVALUATION

In this section, we present the results achieved by our proposal. First, we present the results of the classification block of our methodology, discussing its accomplishments and limitations. Then, we present and discuss and compare the clusters found.

4.1 Trip Classification: Is it work or leisure?

***Text Classification: Trip Advisor as Train and Test*.** Our aim with the ATC block is to apply the model trained with the TripAdvisor dataset, to the Yelp dataset, hence, identifying work and leisure travels. To accomplish so, first, we evaluate the quality of the model trained on TripAdvisor, and then the quality of this model applied to the dataset.

First, Table 3 show each model's effectiveness evaluated in our methodology, presenting each algorithm's Macro-F1, Micro-F1, and Confidence Interval (CI). Even though the RoBERTA approach presents the best average results, all the algorithms are statistically equivalent. These results are referred to the TripAdvisor dataset.

Model	Macro-F1	Micro-F1
Bart	69.1(1.52)	80.86(1.89)
Bert	67.36(1.45)	79.78(1.36)
RoBERTA	70.16(1.57)	82.15(2.11)

Table 3: Models metrics and CI - Train and Test on TripAdvisors data

Based on the results, it can be inferred that the algorithms demonstrate strong predictive capabilities for both classes. The evaluation of Macro-F1 metrics supports the notion that the algorithms achieve favorable classification outcomes for both classes in this dataset.

***Text Classification: Training on TripAdvisor and Predictiing on Yelp*.** As mentioned before, Yelp dataset instances are not labeled, hence, it is not possible to properly evaluate the effectiveness of the ATC. To overcome such a problem, we randomly choose 1095 instances and manually classified them with one reviewer, hence having a ground truth. In the manually labeled dataset, about 72.8% of the instances were classified as leisure visits, while 27.1% was classified as a work visit. Then, we used RoBERTA to classify the selected instances, and we compared the results in terms of Micro and Macro F1. Table 4, shows the results of RoBERTA on the Yelp dataset. Complementary, Table 5, presents the confusion matrix of the model.

Macro-F1	Micro-F1
49.9 %	68.16 %

Table 4: Effectiveness of RoBERTA in the Yelp dataset

It is possible to notice that in general, the model tends to make more mistakes in work-related trips, due to the fact that is the minority class. Besides that, in both datasets, there are reviews made by users that do not clearly specify if the user is making work travel. Illustrating that, we have the following example, which is work-related travel: ‘"I don’t really understand much of the exhibits, but it can be an eye-opener. The free exhibits took us slightly less than 2hrs to complete. The navigation is easy. Premise is clean and spacious.". However, in this particular example, there are no indications suggesting that this trip is work-related. Therefore, as part of our future work, we aim to incorporate additional features that can help us differentiate between leisure and work-related trips. Our intention is to leverage features that are commonly found across various datasets, thereby enabling the application of our proposed model in different scenarios. Thus, in certain instances, relying solely on reviews may not provide sufficient discernment to accurately differentiate trip labels. Consequently, this limitation has an impact on the Macro-F1 results.

	Prediction	
	Leisure	Work
Leisure	656	105
Work	220	40

Table 5: Confusion Matrix

Model Classification: Train and Test with Yelp manually labeled dataset. As an alternative option to ATC block of our methodology, we also proposed a classification method based on the trip’ features, to distinguish between work and leisure travel on Yelp dataset. Our intention with the ATC was to fully label the Yelp dataset, and then, identify the main characteristics of work and leisure trips through the alternative model features’ importance. Nevertheless, this was not possible due to the poor effectiveness of the ATC block. Thus, to construct this alternative model, we used the manually labeled instances.

Our first approach was to use Logistic Regression (LR). To evaluate LR we first tested the premises behind the algorithm, highlighting those that we were not able to respect:

(1) Binary target variable: In our case, leisure or work
(2) Instances independence: In our case, each travel is unique, and our features are not affected by common factors
(3) Sufficient success and failure cases: We have 900 instances with over 200 instances for the minority class.
(4) **Outliers absence:** In our case, there are a few outliers that we cannot remove due to the few instances that we have. In this scenario, the outliers are the people that visit more than one place during the trip. Figure 3, presets the distribution of the features and shows the outliers.
(5) **Multicollinearity absence:** Description variables must have a low correlation between them. In our case, Figure 4, shows that some of the features are highly correlated.
(6) **Description variables must be linear with the prediction probabilities:** In our case, Figure 5 shows that most of the features used in our modeling are not linear with the prediction probability.

Model	Macro-F1	Micro-F1
Random Forest	47.09(2.73)	68.78(4.41)
Dummy Model - Uniform	45.26(3.76)	47.72(2.68)
Dummy Model - Most Frequent	42.17(1.84)	73.02(5.52)

Table 6: Models metrics and CI - Train and Test on Yelp Manually labeled data (95% confidence)

Thus, in our scenario, given that the data do not comply with some premises required by LR, we opt to test Random Forest and the Dummy Classifier (DC). In the DC we use two different strategies to define the prediction values, a **uniform strategy** that randomly assigns a label to an instance, based on the datasets’ original distribution. Also, we use a **most frequent** strategy that defines all instances as the most frequent class in the original dataset, in our scenario leisure. Table 6, shows the results by each approach.

Evaluating the Table, it is possible to notice that the Random Forest classifier is statistically equivalent to the Uniform Dummy Model in terms of Macro-F1, and statistically equivalent to the Most Frequent strategy in terms of Micro-F1. This shows that the approach cannot properly distinguish between work and leisure trips. We believe that this is due to the fact that our prediction variables are based on the average values of the trip made by users. However, as seen in Figure 1, most of the travels have only one check-in. Consequently, it impacts the prediction capabilities of our model, given that most of the features cannot properly describe if a trip is work or leisure.

To confirm this assumption, we perform a Kruskal-Wallis test to compare if the features are statistically different. As the test shows, only three features are statically different in each class (p-value < 0.05), namely, the number of visits to attractions, and the features that show the visit month as May or December. In this scenario, we believe that the difference between the month variables is due to the fact that in December, given the holidays, people tend to travel more in leisure. We compare this assumption by measuring the IC of the percentage of travels along the year. With 95% confidence, 68.9 - 76.76 % of the people make leisure travels along the year, while in December this number goes to 85.39% of the time, clearly showing a temporal impact on the leisure travels.

Figure 6, presents the classes’ distribution for each feature used in our model, and the p-value for the Kruskal-Wallis test. As shown in the Figure, it is possible to notice that the distribution between the variables in each class is very similar for both classes, the main difference is the frequency, given that we have a larger number of instances for the leisure class.

In our last study, we evaluated, for the Random Forest Classifier, the features that have the most effect on the target prediction. For this evaluation, we measure the confidence interval for each feature over the folds with a 95 % confidence. Table 7, shows the features ordered by importance, and if it is statically equivalent. In this scenario, the most important features are the distance to the users' hometown, the check-in time, and the average stars. Even though, the model coefficients are statically different, according to the Kruskal-Wallis test these features are statically equal for both classes. Nevertheless, this can be indicative of the models' poor performance, due to the fact that the model mostly relies on features that have no statistical difference between the classes.

***Summary*.** Answering our **first research question: Is it possible to properly identify which type of trip a traveler is making (work or leisure)?** According to our results, it is possible to predict with good results using a text classification model trained and predicting on TripAdvisor, however, when we tried to apply this model on the Yelp dataset, it had a poor performance, we believe that due to the difference between the social networks, while TripAdvisor is more focused on tourism, Yelp has a more daily characteristic, which enables the users to review not only POI, but also places that are not tourism related.

In our second approach to this task, we manually labeled over 1000 randomly selected Yelp instances, and then, extract features that could characterize the trip made by the user. Again, the models used had a poor performance, due to our data modeling. Due to the users' characteristics, when we group the trips, most of the travels had only one place visited. As most of our features are based on the average of the places visited, we could not model features that properly distinguish between work and leisure. Nevertheless, it was possible to notice a temporal effect on this problem, where most of the trip made in December are leisure related. Thus, as future work, we believe that is possible to explore more of temporal features aiming to achieve better results on the trip purpose task.

4.2 Clustering Analysis

Following our purpose, the second step of our methodology proposes a characterization of tourists, aiming to understand their behavior, and if the taxonomy proposed in other literature works are valid for the dataset that we are using.

For this step, as we had poor performance results in the classification of the trip purpose, we use the Yelp manually labeled dataset, due to the fact that we have a ground truth to remove the work trips. Thus, first, we remove the work trips from the dataset, given that our aim is to evaluate the behavior of tourists. Then, we group the users' trips, having only one instance per user. In the grouping, we sum the one-hot-encode features and use the mean for the other features. After we filter and grouped the users' travels, a total of 415 instances were used to perform the clustering analysis. It is noteworthy that in this step we used the same features that were used in the classification step.

After generating our clustering data, we use K-Means clustering to identify the subgroups within our data. First, we use the silhouette method to identify the optimal cluster amount. In our case, the optimal number was 3. The Figure 2, shows the Krusal-Wallis test evaluating the difference between the features in each cluster. For this visualization, we left out all the month features, due to the fact that they were statistically equal between the clusters. We also show inside each cell if the feature is different within the cluster, or statistically equal (E).

By analyzing Figure 2, it is possible to understand which features made more difference to generate each cluster. First, it is noteworthy that some features do not have an impact on the cluster definition, given that it is statistically equal among all clusters. Nevertheless, it is possible to notice that features (e.g. amount of places, hotel distance, radius gyration, visits to restaurants, hotels and attractions, and distance to the hometown) are different among all clusters, thus, we compare and study the differences between these features, so we can understand tourist behavior. Complementary, Table 8, shows the differences between the main features in each of the clusters. Through the Table, it is possible to understand the users' patterns in each cluster.

The users on the first cluster tend to make visits to places near home (maximum of 30 km radius), visiting attractions and restaurants. As these users tend to visit places that are near their hometown, they usually do not need hotels, and their trips are only visits to one place before their go home, as we established in our trip identification rules.

The users in the second cluster (namely cluster 1), have a similar behavior to the users in the first cluster. The main difference is that these users tend to go to visit places that are more distant, with more than 1000 km far from their home city. As the users in the first cluster, these users do not make check-ins in hotels, this can be justified for several reasons, such as the user does not use Yelp to evaluate their hotels, or the user is staying with some relatives or friends.

Lastly, users in cluster three are the more active type, among all the users evaluated. The users in this cluster, in general, tend to visit more than one place per travel, having a larger radius of gyration, and visiting several restaurants and attractions. As the amount of instances in this cluster is very few (only 12), is harder to assert if this profile is usually in cities near the users' hometown or far, given the large difference between the users within this cluster.

In conclusion, even though we have a small labeled dataset, it is possible to give a taxonomy for the clusters found. The first can be described as more of a stay-near-home cluster, while the second is characterized by persons that are traveling for a far-from-home, and the third cluster is characterized by more active people that tend to visit several different places while traveling.

***Summary*.** Answering our **second research question: What are the main differences between the users' profiles identified by our methodology?** In our evaluation, it was possible to clearly distinguish characteristics in each cluster, while the first cluster is represented by travelers that tend to visit places that are near home, the second cluster is characterized by people that tend visits places that are far from home. Lastly, the third cluster is characterized by people that are more active and visits a lot of different places in travel.

Our **third research question: Does the taxonomy proposed in [2, 13], still holds up when using a properly pre-processed dataset?**, when making such evaluation is hard to compare the data used in our work, and the data used in these approaches. This

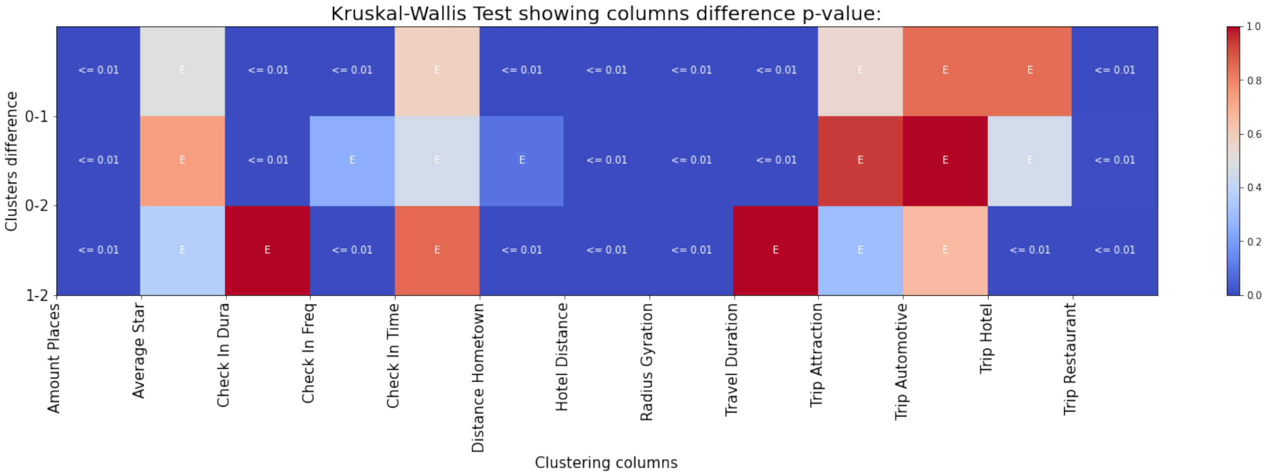

Figure 2: Krusal-Wallis test, showing which features are different between the clusters

is due to the fact that we were not able to scale our proposal to label the complete dataset, thus, the dataset used in the clustering analysis is considerably smaller than the one used in those works. Illustrating that, the work of [13] makes a clustering analysis with over 1 million instances, with users with at least 7 days of the trip. As we considered the same features that were used in their work, the main difference is that our travelers usually make one-day trips, which makes our evaluation much harder, as the data used tends to be sparser. Nevertheless, in the future, we aim to expand the dataset used in this work, hence increasing the number of users in our dataset and the number of clusters.

5 CONCLUSION

In this work, we propose two techniques to define the users' trip purpose, one based on text classification, and one based on classic machine learning approaches. Besides that, we also propose a characterization of tourists, statistically comparing the instances in each cluster and identifying them. Our main goal was to properly characterize tourists, removing from the dataset people that were making work travel.

Our results, in the first block of our proposal, the trip purpose (e.g. work or leisure) classification, were not good, given the difficulty of the task. First, we use state-of-the-art text classification approaches trained in label data from TripAdvisor. This model was then tested in TripAdvisor data and showed promising results. We then, tested the model, in an unlabeled Yelp dataset, hoping that the algorithm would perform well also as in a transfer learning task. To evaluate the algorithm performance on Yelp data, we randomly choose over 1000 instances and manually labeled them, these instances were then classified with our text classification method, however, showed poor performance results. These poor results, lead us to evaluate then another approach to classify the users' travels. In this scenario, we used a Random Forest to train with features extracted from the user trip, and then classify it. One more time, we do not achieve satisfactory results, this time, due to the fact that the users usually only visit one place per trip, and most of the features used in our dataset were based on the average of the features. Thus, most of the instances were equal, making it impossible for the classifier to distinguish between the classes. Hence, answering our first research question, we could not propose a method that properly distinguishes work and leisure travel.

For the second part of our methodology, we proposed a characterization of tourists. As our focus was tourists, and the only properly labeled dataset that contained mobility features that we had was the Yelp dataset that we manually classified, we used this dataset to cluster the users. In this step, we found 3 different clusters with users with different patterns, thus answering our second research question. The first group of users shows a tendency to visit near-home places, while the second places that are far from home (over 1000 km), and the third group was characterized by more active people that visited lots of different places. For our third research question, we aimed to answer if we could extend the current taxonomy proposed given to tourists' characteristics proposed in the literature. However, given that we had a small dataset, we could not expand on the conclusions previously found. Nevertheless, we believe that if we expand the dataset, it will be possible to properly characterize the users of Location-Based Social Networks.

REFERENCES

[1] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* 9, 8 (2020), 1295.

[2] Francisco AmArAl, Teresa TiAgo, and Flávio TiAgo. 2014. User-generated content: tourists' profiles on Tripadvisor. *International Journal of Strategic Innovative Marketing* 1, 3 (2014), 137–145.

[3] Francisco Amaral, Teresa Tiago, Flávio Tiago, and Androniki Kavoura. 2017. Comentários no TripAdvisor: Do que falam os turistas? *Dos Algarves: A Multidisciplinary e-Journal* 2, 26 (2017), 47–67.

[4] Jie Bao, Yu Zheng, David Wilkie, and Mohamed Mokbel. 2015. Recommendations in location-based social networks: a survey. *GeoInformatica* 19, 3 (2015), 525–565.

[5] Punam Bedi, Sumit Kumar Agarwal, Vinita Jindal, et al. 2014. Marst: Multi-agent recommender system for e-tourism using reputation based collaborative filtering. In *International Workshop on Databases in Networked Information Systems*. Springer, 189–201.

[6] Arjun Kumar Bhatia. 2006. *International tourism management.* Sterling Publishers Pvt. Ltd.

[7] Buru Chang, Yookyung Koh, Donghyeon Park, and Jaewoo Kang. 2020. Content-Aware Successive Point-of-Interest Recommendation. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 100–108.

[8] Chen Cheng, Haiqin Yang, Irwin King, and Michael Lyu. 2012. Fused matrix factorization with geographical and social influence in location-based social networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 26.

[9] Erik Cohen. 1972. Toward a sociology of international tourism. *Social research* (1972), 164–182.

[10] Washington Cunha, Felipe Viegas, Celso França, Thierson Rosa, Leonardo Rocha, and Marcos André Gonçalves. 2022. A Comparative Survey of Instance Selection Methods applied to NonNeural and Transformer-Based Text Classification. *Comput. Surveys* (2022).

[11] Washington Cunha, Felipe Viegas, Celso França, Thierson Rosa, Leonardo Rocha, and Marcos André Gonçalves. 2023. An Effective, Efficient, and Scalable Confidence-Based Instance Selection Framework for Transformer-Based Text Classification. *SIGIR* (2023).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Linus W Dietz, Rinita Roy, and Wolfgang Wörndl. 2019. Characterisation of traveller types using check-in data from location-based social networks. In *Information and Communication Technologies in Tourism 2019: Proceedings of the International Conference in Nicosia, Cyprus, January 30–February 1, 2019*. Springer, 15–26.

[14] Gregory Ference, Mao Ye, and Wang-Chien Lee. 2013. Location recommendation for out-of-town users in location-based social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 721–726.

[15] Zachary Friggstad, Sreenivas Gollapudi, Kostas Kollias, Tamas Sarlos, Chaitanya Swamy, and Andrew Tomkins. 2018. Orienteering algorithms for generating travel itineraries. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 180–188.

[16] Benjamin Garner, Corliss Thornton, Anita Luo Pawluk, Roberto Mora Cortez, Wesley Johnston, and Cesar Ayala. 2022. Utilizing text-mining to explore consumer happiness within tourism destinations. *Journal of Business Research* 139 (2022), 1366–1377.

[17] Glauber D Gonçalves, Flavio Figueiredo, Jussara M Almeida, and Marcos A Gonçalves. 2014. Characterizing scholar popularity: a case study in the computer science research community. In *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, 57–66.

[18] Paulene Govender and Venkataraman Sivakumar. 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric pollution research* 11, 1 (2020), 40–56.

[19] Qing Guo, Zhu Sun, Jie Zhang, and Yin-Leng Theng. 2020. An Attentional Recurrent Neural Network for Personalized Next Location Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 83–90.

[20] Qiang Hao, Rui Cai, Changhu Wang, Rong Xiao, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. 2010. Equip tourists with knowledge mined from travelogues. In *Proceedings of the 19th international conference on World wide web*. 401–410.

[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[22] Amir Khatibi, Fabiano Belem, Ana P Silva, Dennis Shasha, Marcos A Goncalves, et al. 2018. Improving tourism prediction models using climate and social media data: A fine-grained approach. In *Twelfth International AAAI Conference on Web and Social Media*.

[23] Jinyoung Kim, Hyungjin Kim, and Jung-hee Ryu. 2009. TripTip: a trip planning service with tag-based recommendation. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. 3467–3472.

[24] Serhan Kotiloglu, Theodoros Lappas, Konstantinos Pelechrinis, and PP Repoussis. 2017. Personalized multi-period tour recommendations. *Tourism Management* 62 (2017), 76–88.

[25] Hyejin Lee and Youngok Kang. 2021. Mining tourists' destinations and preferences through LSTM-based text classification and spatial clustering using Flickr data. *Spatial Information Research* (2021), 1–15.

[26] Antônio HG Leite, Fabrıcio Benevenuto, and Mirella M Moro. 2013. TripTag: Ferramenta de planejamento de viagens baseada em experiências de usuários de redes sociais. In *28 TH BRAZILIAN SYMPOSIUM ON DATABASES*. 37.

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[28] Huayu Li, Yong Ge, Richang Hong, and Hengshu Zhu. 2016. Point-of-interest recommendations: Learning potential check-ins from friends. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 975–984.

[29] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. 2014. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 831–840.

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[31] Yiding Liu, Tuan-Anh Nguyen Pham, Gao Cong, and Quan Yuan. 2017. An experimental evaluation of point-of-interest recommendation in location-based social networks. *Proceedings of the VLDB Endowment* 10, 10 (2017), 1010–1021.

[32] Andrew Ng. 2016. Nuts and bolts of building AI applications using Deep Learning. *NIPS Keynote Talk* (2016).

[33] Olanrewaju Jelili Oyelade, Olufunke O Oladipupo, and Ibidun Christiana Obagbuwa. 2010. Application of k Means Clustering algorithm for prediction of Students Academic Performance. *arXiv preprint arXiv:1002.2425* (2010).

[34] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[35] Philip L Pearce. 2013. *The social psychology of tourist behaviour: International series in experimental social psychology*. Vol. 3. Elsevier.

[36] Lior Rokach and Oded Maimon. 2005. Clustering methods.

[37] Thiago H Silva, Pedro OS Vaz de Melo, Jussara M Almeida, Juliana Salles, and Antonio AF Loureiro. 2014. Revealing the city that we cannot see. *ACM Transactions on Internet Technology (TOIT)* 14, 4 (2014), 1–23.

[38] Artur Starczewski and Adam Krzyżak. 2015. Performance evaluation of the silhouette index. In *Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-18, 2015, Proceedings, Part II 14*. Springer, 49–58.

[39] Haodong Sun, Yanyan Chen, Jianhui Lai, Yang Wang, and Xiaoming Liu. 2021. Identifying tourists and locals by K-means clustering method from mobile phone signaling data. *Journal of Transportation Engineering, Part A: Systems* 147, 10 (2021), 04021070.

[40] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. 2013. Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems*. 374–383.

[41] Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840* (2019).

[42] Budhi S Wibowo and Monica Handayani. 2018. A genetic algorithm for generating travel itinerary recommendation with restaurant selection. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. IEEE, 427–431.

[43] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 325–334.

[44] Haochao Ying, Jian Wu, Guandong Xu, Yanchi Liu, Tingting Liang, Xiao Zhang, and Hui Xiong. 2019. Time-aware metric embedding with asymmetric projection for successive POI recommendation. *World Wide Web* 22, 5 (2019), 2209–2224.

[45] Kyung-Hyan Yoo, Marianna Sigala, and Ulrike Gretzel. 2016. Exploring TripAdvisor. In *Open tourism*. Springer, 239–255.

[46] Chenyi Zhang, Hongwei Liang, and Ke Wang. 2016. Trip recommendation meets real-world constraints: POI availability, diversity, and traveling time uncertainty. *ACM Transactions on Information Systems (TOIS)* 35, 1 (2016), 1–28.

[47] Chenyi Zhang, Hongwei Liang, Ke Wang, and Jianling Sun. 2015. Personalized trip recommendation with poi availability and uncertain traveling time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 911–920.

[48] Jia-Dong Zhang and Chi-Yin Chow. 2015. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 443–452.

[49] Jia-Dong Zhang, Chi-Yin Chow, and Yanhua Li. 2014. Lore: Exploiting sequential influence for location recommendations. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 103–112.

[50] Shenglin Zhao, Irwin King, and Michael R Lyu. 2016. A survey of point-of-interest recommendation in location-based social networks. *arXiv preprint arXiv:1607.00647* (2016).

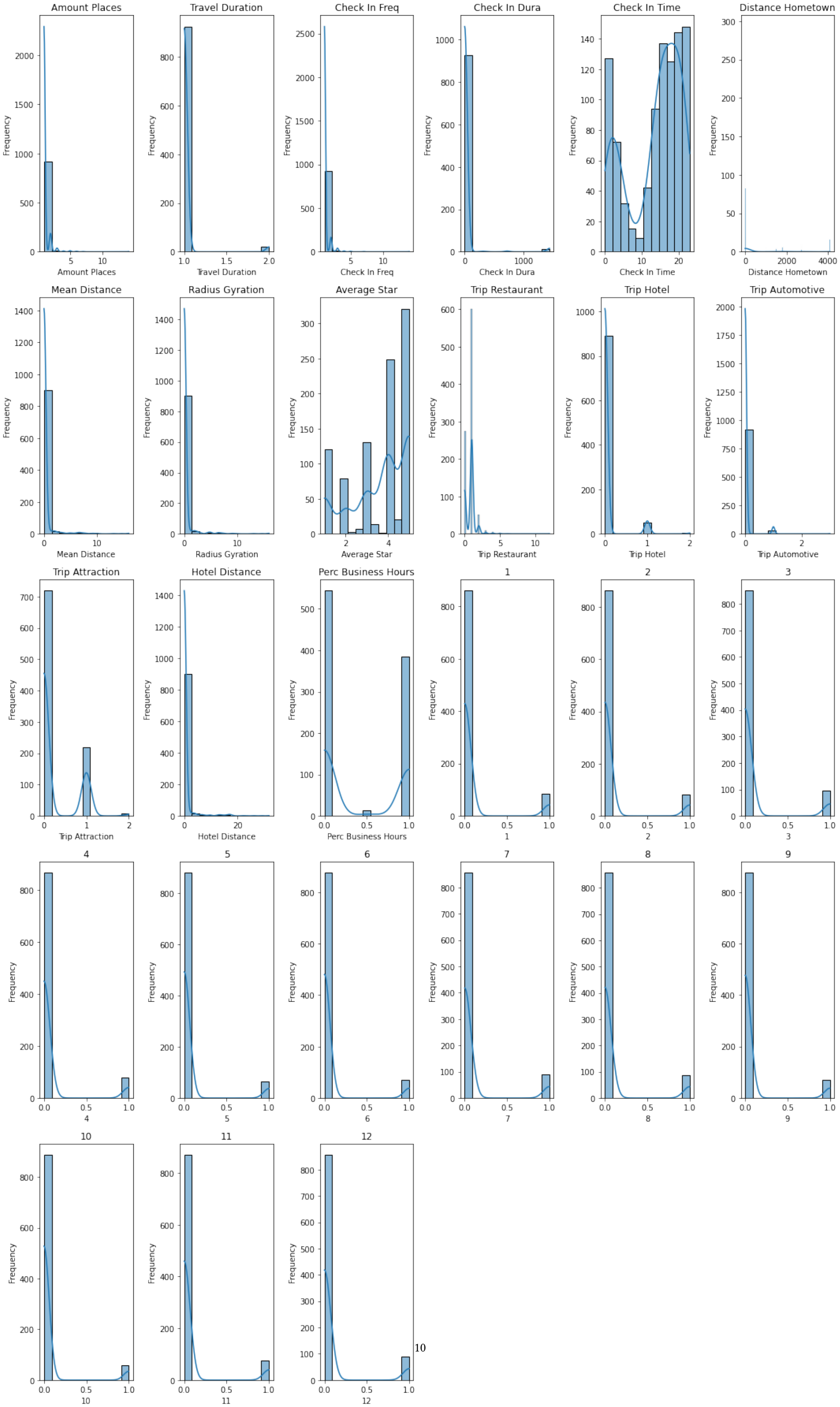

Figure 3: Distribution of the features used in the classification model

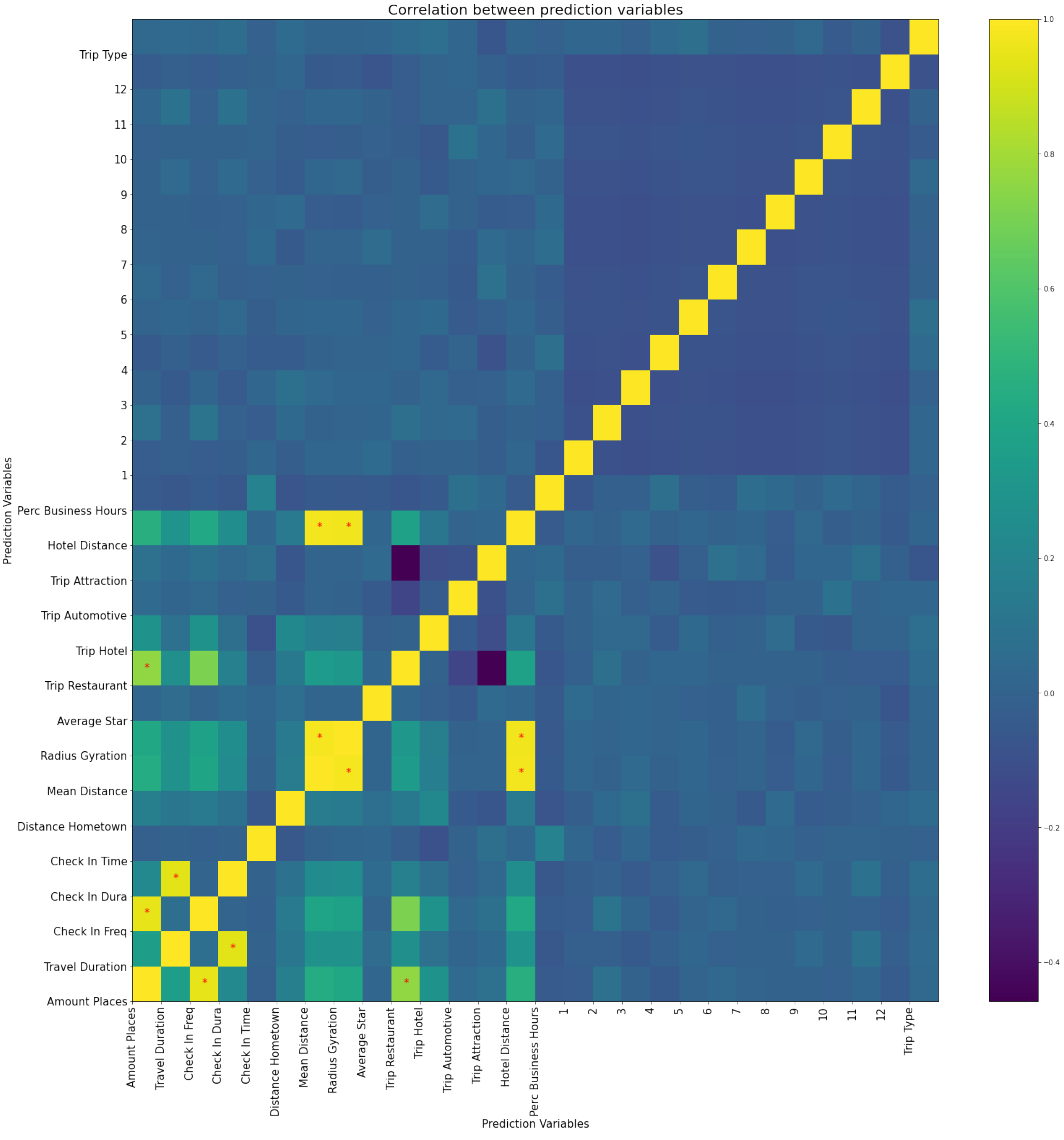

Figure 4: Correlation between the features used in the classification model - In red we highlight the variables with a correlation above 0.75

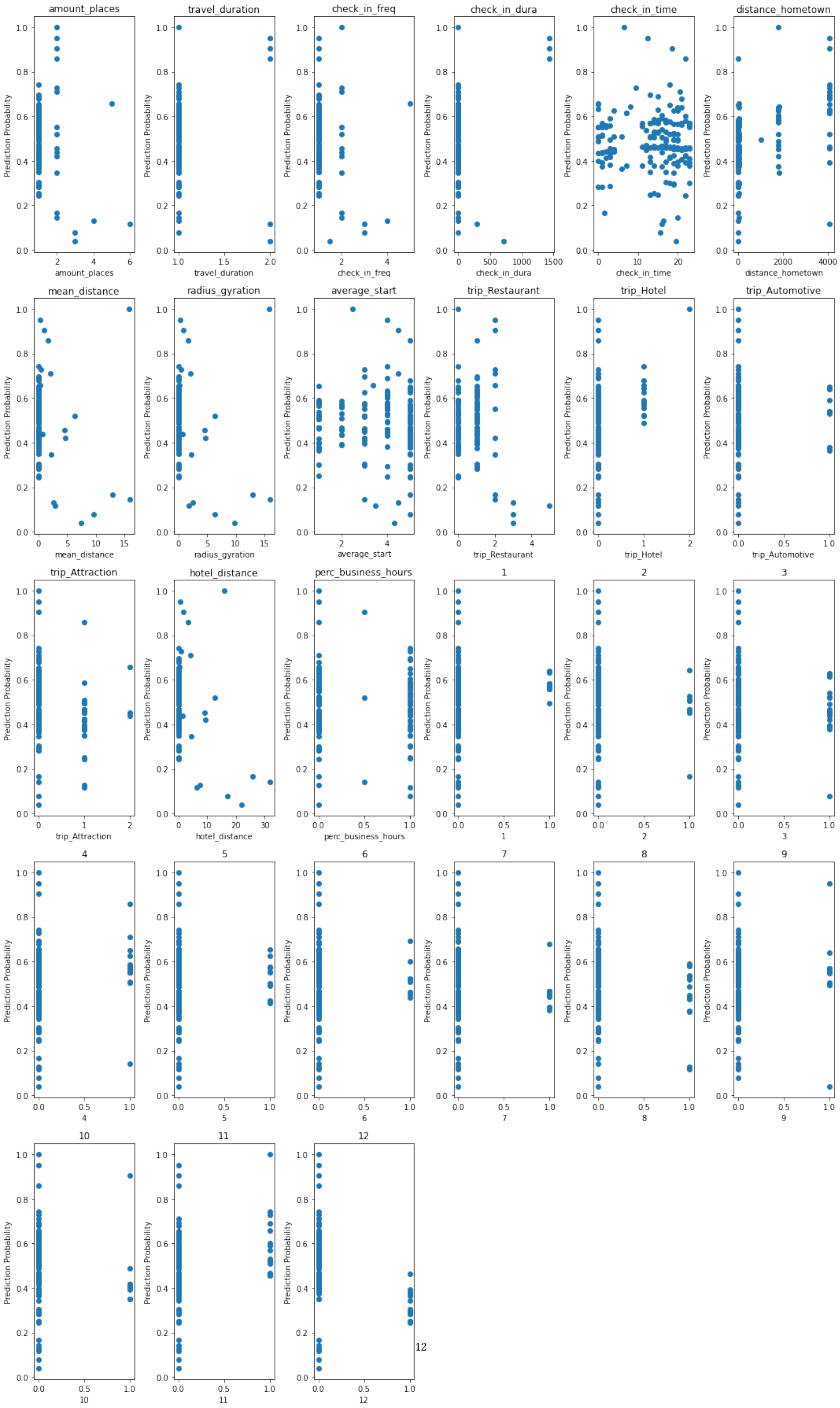

Figure 5: Feature values x LR Prediction Probability

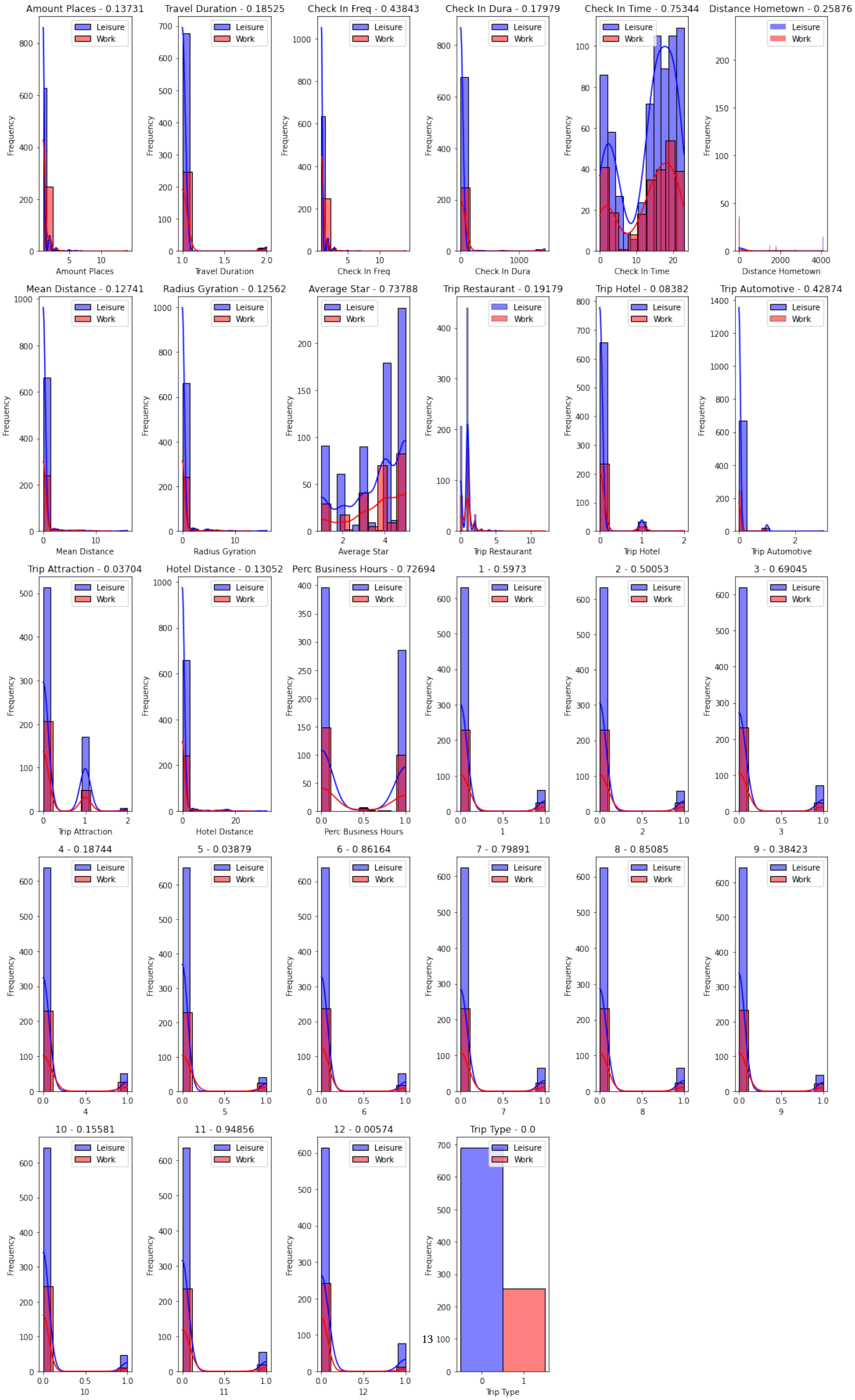

Figure 6: Distribution of the features used in the classification model for each class

Descriptive Feature	Lower Importance	Upper Importance	Equivalent
Distance Hometown	0.634637	0.644372	-
Check-In Time	0.384834	0.396083	False
Average Star	0.213503	0.219180	False
Perc Business Hours	0.061044	0.064784	False
Trip Restaurant	0.050998	0.056624	False
Trip Attraction	0.048424	0.051484	True
12	0.040372	0.043487	False
1	0.035604	0.041861	True
3	0.036017	0.040028	True
2	0.034461	0.038088	True
7	0.035460	0.036874	True
10	0.031581	0.036671	True
8	0.035019	0.036428	True
4	0.031324	0.036063	True
6	0.031283	0.035933	True
5	0.031187	0.034293	True
11	0.029859	0.034083	True
9	0.031792	0.033770	True
Hotel Distance	0.025600	0.031158	False
Radius Gyration	0.027639	0.030060	True
Mean Distance	0.027187	0.029923	True
Trip Hotel	0.021634	0.023209	False
Trip Automotive	0.015443	0.018882	False
Check-In Freq	0.011900	0.014481	False
Amount Places	0.012120	0.013513	True
Check-In Duration	0.004449	0.008283	False
Travel Duration	0.003080	0.003934	False

Table 7: Random Forest Classifier Feature Importance - The Last column shows if the feature is statistically equivalent to the feature above it

	Amount Places	Hotel Distance	Radius Gyration	Trip Restaurant	Trip Hotel	Trip Attraction	Distance Hometown	Cluster
count	66	66	66	66	66	66	66	1
mean	1.303030	1.193009	0.641313	1.469697	0.090909	0.257576	1874.219071	1
std	0.825424	3.253107	2.177190	2.213384	0.419790	1.193709	1054.610426	1
min	1	0	0	0	0	0	210.804640	1
25%	1	0	0	0	0	0	1174.758341	1
50%	1	0	0	0	0	0	1800.278275	1
75%	1.191667	0.133247	0.066624	2	0	0	1989.997625	1
max	7	16.226987	15.844767	9	2	9	4093.476620	1
count	337	337	337	337	337	337	337	0
mean	1.078380	0.281668	0.118860	0.632047	0.005935	0.145401	10.583655	0
std	0.352411	1.712184	0.688077	1.660441	0.108947	0.808944	7.102759	0
min	1	0	0	0	0	0	1.720854	0
25%	1	0	0	0	0	0	4.821060	0
50%	1	0	0	0	0	0	8.264845	0
75%	1	0	0	0	0	0	14.320972	0
max	5	17.219940	6.198907	16	2	10	30.666363	0
count	12	12	12	12	12	12	12	2
mean	2.666667	5.129116	2.202004	3	0	0.166667	1442.383600	2
std	1.522558	6.242058	2.786647	1.906925	0	0.577350	1747.033176	2
min	1.333333	0.073314	0.026750	0	0	0	5.234759	2
25%	1.458333	0.483635	0.240299	2	0	0	6.136254	2
50%	2	4.755409	1.771907	3	0	0	692.156910	2
75%	3.125000	6.531515	2.731671	4.250000	0	0	2383.742854	2
max	6	22.076702	9.791943	6	0	2	4087.076563	2

Table 8: Clusters' main features summary