Master Thesis

Crowdsourced Product Descriptions and Price Estimations

Steve Aschwanden

Dammstrasse 4

CH-2540 Grenchen

steve.aschwanden@students.unibe.ch

05-480-686

Supervisor

Dr. Gianluca Demartini

C302, Bd de Pérolles 90

CH-1700 Fribourg

demartini@exascale.info

Grenchen, July 7, 2014

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all the principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Steve Aschwanden, 05-480-686

Grenchen; July 7, 2014:

_____

(Signature)

Acknowledgements

First of all, I thank Dr. Gianluca Demartini for the possibility to write my thesis based on a self-defined topic, and for the support and the ideas he was giving me during this time.

Furthermore, I express gratitude to the eXascale Infolab[1] for giving me the opportunity to execute my experiments and for the helpful remarks after the mid-term presentation.

I thank my fellow student Marcel Würsch for the cooperation over the three years of study and for the tips and inputs during the thesis.

I thank my family which believe always in me. A full-time study would not have been possible without you.

---

[1] http://www.exascale.info

Abstract

The creation of auctions for the online marketplace eBay is time consuming and repetitive. The first step for selling an item is to take pictures of it. To complete the auction, the user has to provide a title, description, category, and other default parameters. One of the most important steps is the definition of a starting price.

Crowdsourcing is used to generate the required information for a complete auction based on several images. The complex task is split into multiple subtasks. The thesis presents a pure and a hybrid crowdsourcing approach. Different experiments were made to investigate the behaviour of the crowd.

A promised commission for successful auctions has the biggest influence on the quality of the workers. Workers are a group of people which is responsible to solve tasks on crowdsourcing platforms. A majority favours the results of this experiment over the descriptions of the real online auction. The workers did the most accurate price predictions if the actual market price of the items have been provided.

The results of the executed experiments show the potential of the crowd. If all the strength of the single variations will be combined and the task design improved slightly, then the generated contents can be used to create real auctions on eBay in the future.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

eBay Inc.[1] is one of the world's largest online marketplaces and reported 128 million active users worldwide during the last quarter of the year 2013. Online auction platforms make consumer-to-consumer transactions possible. The seller can present articles by uploading pictures and characterise them by writing proper descriptions. The creation of an auction is time consuming and needs a lot of investigations. For example, search for descriptions on the internet or find a selling prices for the same or similar article. In 2005, Jeff Howe and Mark Robinson created a term called 'Crowdsourcing' which is a combination of the words crowd and outsourcing. The idea behind the term is to outsource different tasks, which are difficult to solve by machines, to the crowd. To reduce the costs of collecting information for an article to sell on an auction platform, tasks will be created and outsourced to the crowd. Amazon Mechanical Turk[2], short MTurk, is a crowdsourcing marketplace which enables requesters to publish human intelligence tasks (HITs). The workers can solve these tasks and earn money for their work.

1.1 Statement of the Problem

The first step of creating an online auction is to take pictures of the item. This helps the buyers to get information about the state and quality of the article. After that, the item needs a short and clear description, some properties (category, state) and a starting bid. If the seller wants to create a lot of different auctions, the whole procedure is time consuming and boring. A price estimation of an article can be difficult because the background knowledge is missing and other auctions to compare are not available at any time. Machines are not able to solve all these steps by them self because the spectrum of the articles is huge and image processing methods are not capable to classify all of them correctly. To get all the needed parts of an online auction, a human powered approach is necessary. Crowdsourcing platforms provide the possibility to solve tasks which are difficult to handle for a computer.

---

[1]http://www.ebay.com
[2]http://www.mturk.com

1.2 Existing Research

Some similar existing research projects will be illustrated in this section:

1.2.1 Crowdsourcing

The idea of the thesis is similar to a project called "PlatMate" [20] where workers analyse the content of food photographs. The processing pipeline consists of three major steps and put out the calorie values of every ingredient on the picture. All steps were performed by workers of a crowdsourcing platform. The accuracy of the calorie estimations of the system was almost as good as estimations from different trained experts.

1.2.2 Price Estimation

The vision of predicting the end price of online auctions is not new. People from the Accenture Technology Labs[3] tried to do this in 2005 and published some surprising results [10]. They collected 1'700 auctions of a specific item during a two-month period to form a training and test set. The end prices of the ground truth are additionally converted to a price class (10% of the average price) to perform classification algorithms. The accuracy of the classifiers are higher than 70%.

1.3 Goals and Objectives

The thesis has the following goals and their corresponding objectives:

- **Collect auction item properties by the crowd.**

  - Analyse the composition of an auction item on eBay and select the parts which can be crowdsourced.

  - Form a ground truth including multiple auctions created by real online auction platform users by using the eBay API (Application Programming Interface).

  - Study literature which covers similar crowdsourcing problems.

  - Design and publish tasks on Amazon Mechanical Turk to gather data from the crowd.

  - Evaluate the quality of the generated contents.

- **Vary the design of the tasks and investigate the behaviour of the workers.**

  - Find parameters for the HITs.

  - Analyse the influence on the performance of the workers.

- **Try to improve the initial solution by implementing a hybrid approach.**

  - Search for image processing or machine learning methods which can simplify and/or support a human intelligence task.

  - Implement the methods and adapt the design of the tasks.

  - Compare the results with the first implementation.

---

[3]http://www.accenture.com

1.4   Organisation

The thesis report is organised in multiple chapters. At the beginning of the document, the eBay marketplace and the corresponding API will be investigated (Chapter 2). Then, the theoretical background knowledge about crowdsourcing is summarised (Chapter 3). The learned theory is used to build up a workflow to generate online auction contents (Chapter 4). The observations of the executed experiments are concluded in the next chapter (Chapter 5). Results of the thesis are discussed in the next-to-last chapter of the report (Chapter 6). A few ideas for improvements and a summary of the pros and cons of the implemented approach are part of the last chapter (Chapter 7). All the ground truth items, some plots and tables which have not found a place in the report are listed in the appendix section.

Chapter 2

eBay Online Marketplace

2.1 History

eBay was founded 1995 in San Jose (CA) as AuctionWeb by Pierre Omidyar. One year later, eBay bought a third-party licence from Electronic Travel Auction to sell plane tickets and other travelling stuff. During the year 1996, over 200'000 auctions were available on the website. At the beginning of 1997 the number of auctions exploded (about 2 million articles). In the same year the company got their well-known name eBay and received 6.7 million dollars from the venture capital firm Benchmark Capital. The company went public on the stock exchange on September 21, 1998 and the share price increased from 18 to 53.5 dollars on the first day of trading. Four years later the growth continued and eBay bought the online money transfer service PayPal[1]. eBay expanded worldwide in early 2008, had hundred million of registered users and 15'000 employees. Today, the firm is one of the world's largest online marketplaces. During the fourth quarter of the year 2013 about 128 million active users were reported. A cell phone was sold every 4 seconds, a pair of shoes every 2 seconds and a Ford[2] Mustang every 55 minutes.

2.2 Auction Item Composition

Every eBay user has the possibility to create auctions for different kind of items. To present the article, the seller has to provide accurate information about it. The standard eBay auction consists of the following fields:

- **Title:** The title of the item is limited to 80 characters. The sellers should use descriptive keywords to clearly and accurately convey what they are selling.

- **Description:** The description is the opportunity to provide the buyers with more information about the item.

---

[1]http://www.paypal.com
[2]http://www.ford.com

- **Category:** An item can have multiple predefined categories. eBay provides a list of categories which the sellers can use to choose the most appropriate ones.

- **Condition:** The condition of the item is dependent on the selected category. eBay provides different condition schemas. For clothing items, the seller can select between 'New with tags', 'New without tags', 'New with defects' or 'Pre-owned'. For other categories like books, other condition values are present: 'Brand new', 'Like new', 'Very good', 'Good', 'Acceptable'.

- **Pictures:** To visualise the item, the auction creator can upload up to twelve pictures. The first image is important because it appears next to the item's title in the search result. The pictures will be stored for 90 days on the eBay servers.

- **Shipping costs:** The seller has to tell the future buyers how much shipping will cost. There are three possibilities:

    - *Free shipping*

    - *Flat shipping*: Same cost to all buyers

    - *Shipping rate tables*: eBay calculates the cost for every individual buyer dependent on the location

- **Duration:** An auction can have a duration of 1, 3, 5, 7 or 10 days. If the item has a fixed price, the auction is finished if a buyer is willing to pay this price.

- **Pricing:** The seller can select a starting price, then the bidding will start at this price. A 'Buy it now' option is also available. The buyer can skip the bidding process.

- **Payment:** The seller has to select the desired payment method like 'PayPal' or 'Payment upon pickup'.

2.3 APIs

eBay provides multiple APIs for developing third party applications. This allows developers to search for auctions or create listings over the XML format. Three main interfaces are available:
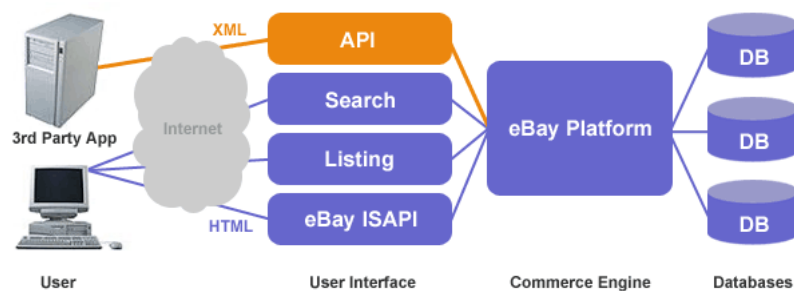
Figure 2.1: eBay API overview

2.3.1   Trading API

Developers use the Trading API to build applications such as selling and post-sales management applications, manage user information, and initiate the item purchase flow on eBay. The API is available in .NET, Java, PHP and Python.

2.3.2   Shopping API

The Shopping API provides a search engine for user information, popular items and reviews. The API is available in PHP and Python. Example calls for this API are:

- *findProducts()*: Search for products by keywords or the ProductId.

- *GetSingleItem()*: Buyer specific view of an item.

- *GetUserProfile()*: Get the user profile and feedback information.

2.3.3   Finding API

The Finding API provides access to the next generation search capabilities of the eBay platform. The developer can search and browse for items based on keyword queries, categories or images. The API is available in .NET, Java and Python. Example calls for the API are:

- *findCompletedItems()*: Find items which are listed as completed or no longer available on eBay.

- *findItemsByCategory()*: Find items in a specific category.

- *findItemsByImage()*: Find items which have a close resemblance to a given image. This call is restricted to items listed in Clothing, Shoes & Accessories category only.

2.3.4   Example

The following listing in Python illustrates the functionality of the Finding API. The developer has to register to the eBay developers program[3] first. After that, an application ID can be created. This is necessary to get access to the eBay databases. A functioning Python environment and the additional eBay Python SDK are requirements to successfully execute the example:

```
1  from ebaysdk.finding import Connection as Finding
2  from ebaysdk.exception import ConnectionError
3  import json
4
5  try:
6      api = Finding(appid='Universi-3c25-4b4e-b3e6-8c2568808b12')
7      api.execute('findCompletedItems', {
8          'keywords': 'ford mustang',
9          'itemFilter': [
10             {'name': 'ListingType',
11              'value': 'Auction'},
12             {'name': 'Currency',
13              'value': 'USD'},
14             {'name': 'SoldItemsOnly',
```

---

[3]http://developer.ebay.com

```
15              'value': 'true'},
16          ],
17          'sortOrder': 'StartTimeNewest',
18          })
19      response = json.loads(api.response_json())
20
21      print response['searchResult']['item'][0]
22
23  except ConnectionError as e:
24      raise e
```

Listing 2.1: eBay Finding API example

The initialisation of the application is done in line 6. A correct application ID is required. Then, the API call *findCompletedItems()* is executed with some keywords and filter options. Only the newest auctions with at least one bidder and a payment in US dollars will be returned. The function *response_json()* (Line 19) returns the first 100 items by default. At the end, the first result will be printed out to the console. Here is a shorter simplified version with the most important fields of the output:

Name	Value
itemId	281273507096
title	2014 Hot Wheels Super Treasure Hunt 71 Mustang Mach 1
categoryName	Diecast-Modern Manufacture
shippingType	Free
currentPrice	18.5 USD
bidCount	1
paymentMethod	PayPal
conditionDisplayName	New
startTime	2014-02-25T04:32:17.000Z
endTime	2014-02-25T05:27:14.000Z

Table 2.1: eBay Finding API example output

Chapter 3

Crowdsourcing

3.1 Introduction

In 2005, Jeff Howe and Mark Robinson created the term 'Crowdsourcing' after a discussion about how businesses can outsource their work to individuals over the internet. There exist multiple definitions in the literature. Enrique Estellés-Arolas and Fernando González Ladrón-de-Guevara analysed over 40 definitions of crowdsourcing and developed a new integrating definition [9]:

"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken."

3.2 Platforms

3.2.1 Amazon Mechanical Turk

The project was introduced in 2005 and is part of the Amazon Web Services[1]. Requesters can post tasks known as HITs (Human Intelligence Tasks) which can be solved by workers (Amazon uses another term: Turkers). MTurk provides a web-based user interface and a couple of APIs in different programming languages (.NET, Java, Python, PHP, Perl, Ruby) to manage tasks. The first action of the requester is to create a HIT consisting of mandatory fields:

---

[1]http://aws.amazon.com

- **Title:** The requester must describe the idea of the HIT in at most 128 characters.

- **Description:** A more detailed description of the task which cannot be longer than 2'000 characters.

- **Question:** Every task has to contain questions to collect information from the crowd. The requester can decide between three question data structures.

  - *QuestionForm*: The simplest form to create questions in a HIT. MTurk uses a special XML language to define tasks which has some restrictions. For example, JavaScript and CSS are not allowed.

  - *ExternalQuestion*: MTurk will display a requester defined external webpage and the answers to the questions will be collected on the external website and send back to MTurk. This question data structure is used to overcome some restrictions of the platform like using JavaScript or to display CSS defined content.

  - *HTMLQuestion*: This structure is a mixture between QuestionForm and ExternalQuestion. The requester has not to host an external website to provide a HTML based form.

- **Reward:** If the workers will successfully completing the HIT, then they will receive a predefined amount of money from the requester.

- **Assignment duration in seconds:** The time in which the workers have to complete the task after they have accepted it. The time has to be between 30 seconds and one year.

- **Lifetime in seconds:** The lifetime of a HIT defines the amount of time a task is acceptable for the workers. After the time elapsed, the HIT will no longer appear in the search results.

and some important, optional fields:

- **Keywords:** Comma separated keywords which describe the task (max. 2'000 characters).

- **Max assignments:** Number of times a HIT can be completed. The default values is one.

- **Qualification requirement:** Requesters can define requirements to process a task for the workers. For example, only workers who have more than 100 approved assignments can start working on a requesters HIT.

After the tasks are designed, the requesters have to test them on the Amazon Mechanical Turk Developer Sandbox platform which is a simulated environment. If the requester is happy with the appearance of the HIT, the task can be published on the productive MTurk platform. Turkers have now the possibility to accept the HITs and complete the assignments until the lifetime is expired. After the HIT is completed, the requesters can take a look at the results and have to decide if they want to accept or reject the work. The workers will receive the predefined amount of money only for an accepted task.

3.2.2   Crowdflower

A platform for large-scale data projects was founded in 2007. Crowdflower[2] has over 50 labor channel partners, Amazon Mechanical Turk for example, where the created tasks are published.

The partner websites or communities are responsible to manage the registration and payment of their workers. The company offers enterprise solutions and enables a higher degree of quality control. 'Gold standard data' (cf. 3.6.2, page 22) and 'Peer review' are two provided quality control techniques. 'Peer review' gives the requesters the chance to improve the data by a second pass. A workflow management tool helps to link different jobs together. At the time of writing these lines, over one billion tasks are completed by workers domiciled in 208 different countries. Big companies like eBay use the Crowdflower service for their projects [7]. Over the past years, the company has completed over 15 projects. The improvement of the product categorisation algorithm was one of them.

3.3 Patterns

This section presents two probed ways to get useful information from the crowd.

3.3.1 Find-Fix-Verify

The Find-Fix-Verify pattern was introduced by the Soylent paper [3]. The pattern divides the overall task into three stages. During the Find stage, the workers will identify patches of work done by the crowd or create new patches. For example, the workers have to select a sentence which seems to be incorrect and will need further investigations during the Fix phase. Some workers will revise the identified patches and try to provide alternatives. The last step of the pattern will present the generated alternatives during the Fix stage to a few new workers in a randomized order. The answer with the most votes (plurality voting) will be used to replace the identified patch during the first phase. The creators of the new suggestions will be suspended so that they cannot vote for their own input.

To illustrate the meaning of the Find-Fix-Verify pattern, the implementation of Soylent will be discussed (Figure 3.1). The approach begins by splitting a text into paragraphs. During the Find stage, the workers have to identify candidate areas for shortening in each paragraph. If a certain number of workers have selected the same area, then this patch goes to the next stage. Every worker in the Fix stage has to present a shorter version of the identified patch if possible. They also have the possibility to say that the text cannot be reduced. During the last step, the crowd has to select rewrites which have spelling, style, or grammar problems or change the meaning of the sentence significantly. At the end, they remove these patches by a majority voting.

3.3.2 Iterative

Most of the published assignments on MTurk are independent, parallel tasks. However, also iterative sequential tasks can be useful. The authors of the TurKit paper [18] implemented a tool which make iterative tasks possible. They developed an example application for creating an image description (Figure 3.2). During the first iteration, the worker will contribute the initial description of the provided image. The next iteration will show the initial description and a request to improve it. A few workers will evaluate the extension of the description by voting. If the extended description does not receive enough votes, then the iteration will be ignored. The final description is generated after a fixed number of iterations. To make the iterative solution possible, the crash-an-rerun programming model was introduced by the authors of the paper. This
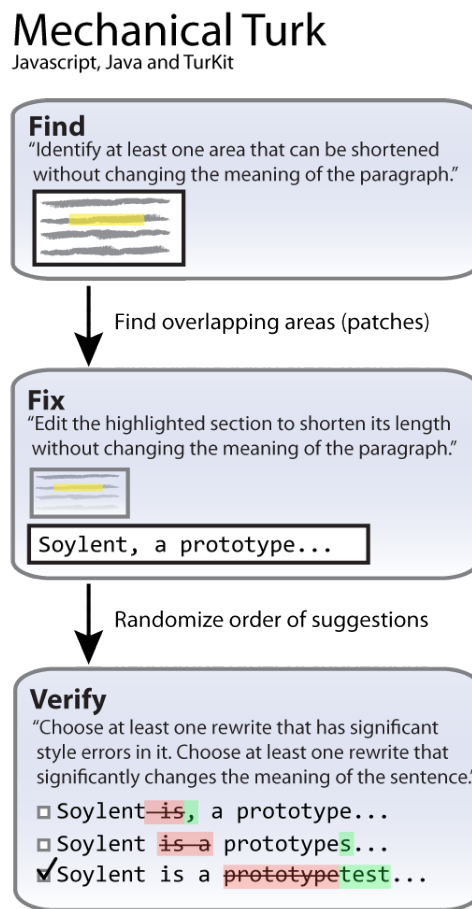
Figure 3.1: Soylent Fix-Find-Verify pattern

model allows a script to be re-executed after a crash without generating costly side-effects. This means, if there is a crash during the second iteration of an iterative problem, the first iteration will be skipped after re-running the script. TurKit is able to persist the state of the program and will never repeat successfully completed tasks. This is helpful for prototyping algorithms.

3.4   Design

If requesters want to create new HITs, then they have to consider some design guidelines [1, 2]:

- **Be as specific as possible in the instructions:** If the requesters ask the workers "Is a Ford Mustang a sports car?", then this is not the same as they ask them "Can a Ford Mustang accelerate from 0 to 100 km/h in 3 seconds or less?" because the second one is clearer and more precise. Sometimes it is useful to hire a technical writer for phrasing task instructions.

- **Instructions have to be easy to read:** Instructions should be split into multiple subtasks and presented as a bulleted list.

- **Provide examples:** The best way to present the idea of a task is to show one or multiple examples. For example, this can help to avoid uncertainties if the instructions are
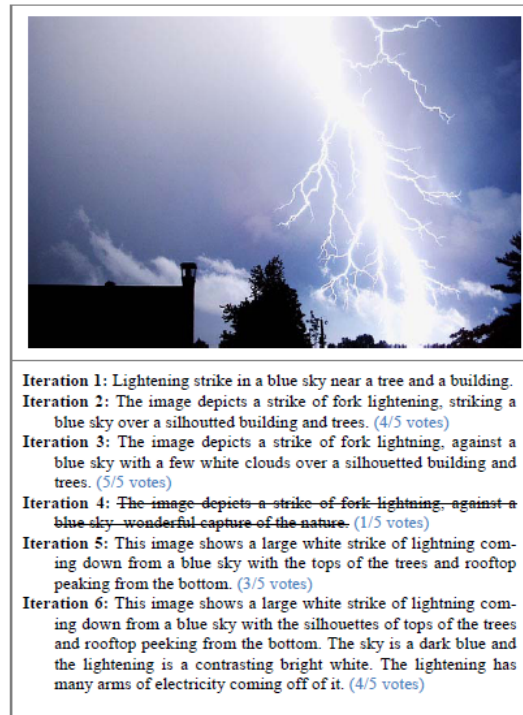
Figure 3.2: Iterative image description created by TurKit

misinterpreted or the workers have wrong expectations.

- **Mention what will not be accepted:** If a worker has to write a paragraph about an encyclopaedia article, the requester can allude in the instructions that copying contents from other website are prohibited.

- **Tell the workers which tools they should use.**

- **Give the workers the possibility to write down a feedback about the task:** This is important to improve the design of the tasks, or can help to detect spammers.

- **Iterative and incremental development of tasks:** The first draft of a task will never be perfect. With the feedbacks and results of the previous iterations, the next one will contain improvements which should avoid foregoing mistakes or design failures.

3.5 Hybrid

A lot of information systems use a hybrid crowdsourcing technology. The combination of human intelligence and machine algorithms can lead to powerful information systems which cannot be realised by a pure machine approach. In most cases, the crowd is responsible to verify the created content of machine algorithms or to generate input data for them. A closer look at the CrowdSearch [25] project helps to illustrate the idea of hybrid systems. The developers implemented an image search system for cell phones. First, the system uses an automated image search to generate a set of candidate pictures. These are packed into multiple identical tasks for validation by humans and published on Amazon Mechanical Turk (Figure 3.3). A simple majority voting is used to

Query Image   **Candidate Images**   **Duplicate Validation Tasks**

Results	1	2	3	4	5
C1	Yes	Yes	Yes	Yes	Yes

Results	1	2	3	4	5
C2	Yes	No	Yes	Yes	Yes

Results	1	2	3	4	5
C3	No	No	No	No	No

Results	1	2	3	4	5
C4	Yes	No	No	Yes	Yes
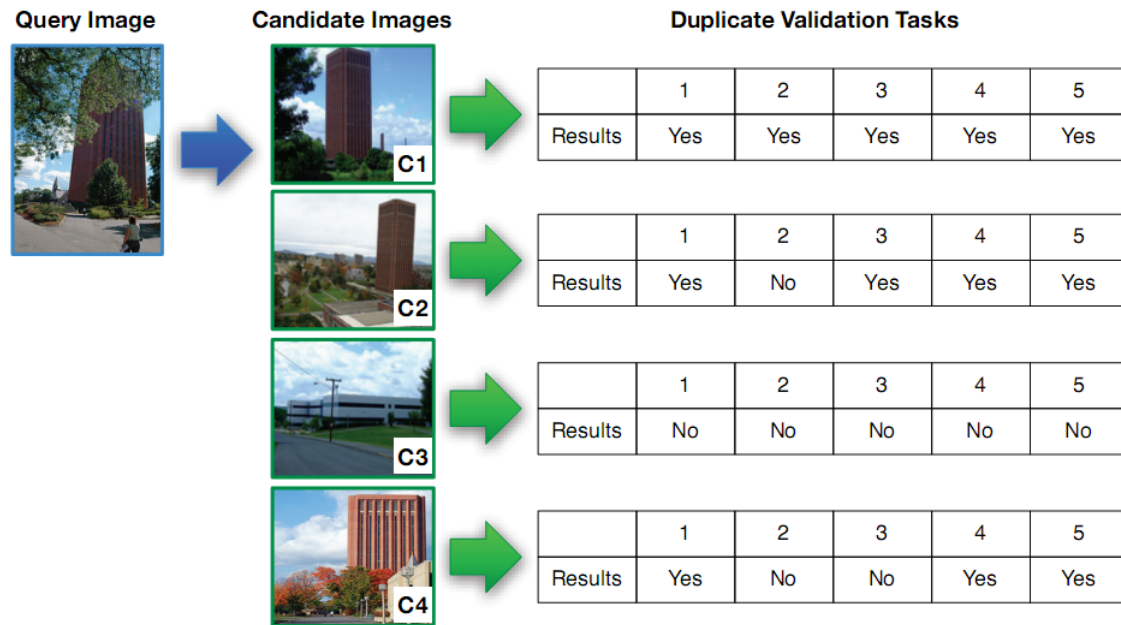
Figure 3.3: CrowdSearch hybrid image search approach

eliminating errors. After the validation of the results, the resulting image will be presented to the user. The drawbacks of such systems are that the hybrid approach generates additional costs for involving humans and the delay between publishing the tasks and receiving the corresponding results. The users of CrowdSearch can define a deadline before they query an image and the system will always return a result after the time is expired, irrespective of whether the crowdsourced tasks are completed or not.

3.6 Quality Control

Determination of the quality of completed tasks by the crowd is very important. Workers can be lazy or spammers who want to earn money for free or a minimal amount of work. To evaluate the performance of a single worker, several techniques are available.

3.6.1 Majority Voting

To reduce the errors of single workers, majority voting can be used. If a majority has the same answer to a question, the requester can assume that the answer is correct. To break ties, an expert is necessary.

3.6.2 Honey Pots

The requesters include trap questions where they know the correct answer. If the answer of a single worker is incorrect, the requester can exclude the results or reject the task. However, it is not always possible to generate honey pots.

3.6.3   Qualification Test

MTurk provides the possibility to include a qualification test at the beginning of tasks. The worker has to pass the test to has access to the real tasks and the resulting rewards. The results of the test can be compared to an answer key automatically or by the requesters themselves. The additional effort and the determent of some workers are drawbacks of this procedure.

3.7   Workflow

A workflow is a set of tasks which are interconnected and easier to solve by the crowd. The output of a single subtask will be used for one or multiple subsequent subtasks. The output of the last element of the flow is the result of the entire complex task. There exists a lot of literature which covers the problematic of finding and interconnecting subtasks:

The process of decomposing complex tasks into simpler ones is not always easy and needs a lot of clarifications. The developers of the Turkomatic [17] tool had an innovative idea and sourced the workflow decomposition out to the crowd. The workers have to decide how the final workflow should look like and what are the belonging tasks. The system consists of two major parts. The meta-workflow is used to design and execute workflows by applying the price-divide-solve (PDS) procedure. The workers have to recursively divide the complex task into smaller ones until they are simple enough. After this step, the workers will solve the generated tasks and other workers are asked to check the solutions. At the end, the results are combined into a cohesive answer. The second part of the Turkomatic system allows a visualisation of the created workflows and an edit function to manually adapt the crowdsourced results.

Another idea was pursued by the developers of CrowdForge [16]. They designed a framework to create a workflow by using several partition, map and reduce steps. The partition step splits a larger task into smaller subtasks, the map step lets one or more workers process a specified task. The results of the workers are merged into a single output during the reduce step. The workers should write an encyclopaedia article about a given topic (Figure 3.4), for example. The authors of the paper solved this problem by the presented partition/map/reduce steps. First, the partition step asks the workers to create an outline of the article by defining section headings (e.g. "History", "Geography"). During the map phase, multiple workers are asked to provide a single fact about the section (e.g. "The Empire State Building celebrated its 75th Anniversary on May 1, 2006" if it is an encyclopaedia article about "New York" and the section heading is "Attractions"). The workers have to piece the collected facts together to a completed paragraph during the reduction step.

The CrowdForge prototype is written in Python using the Django[3] web framework and boto[4], an interface to the Amazon Web Services which is available in Python. The user can define complex flows by creating HIT templates (which can be either a partition, map or reduce task) and dependencies between the templates. Flows are implemented as Python classes. The prototype is also responsible for the sequential coordination between the HITs (including data transfer). Multiple independent flows can be executed simultaneously. One of the limitations is that CrowdForge does not support iteration or recursion. The further development of the project was suspended in

---

[3]https://www.djangoproject.com
[4]https://github.com/boto/boto
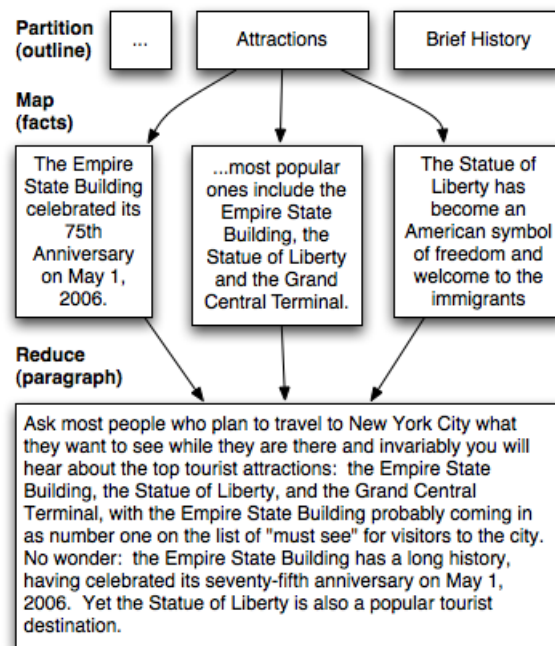
Figure 3.4: CrowdForge example workflow

2011.

The same crew developed CrowdWeaver [15] which is an advancement of the CrowdForge project. They use CrowdFlower, another crowdsourcing platform, instead of Amazon Mechanical Turk. On CrowdFlower, the requesters can create tasks on multiple markets (including MTurk). Flows can be created visually and does not assume any programming skills. Another feature is the tracking and notification of crowd factors, for example latency or price.

3.8 Incentives

There are multiple aspects which motivate users to contribute their human power and knowledge. Some of them are described in the following lines.

3.8.1 Gamification

The ESP game [22] makes the labelling of any kind of images in the web possible. There are no guidelines to provide images and no computer vision method exists which can handle the diversity of all images. Search engines are dependent on accurate image descriptions to represent relevant results. Therefore, another approach was introduced by the article. An online, web-based game was developed to attract workers. Two players are randomly assigned to label the same image simultaneously. There is no possibility to communicate with the game partner. Each player has to guess the description of the image independently without uses the 'Taboo words'. These words are evaluated by a prior round and will be ignored for the actual turn. If there is a match between both players, the score will be increased and another image description is detected. The discovered word will only be taken as a valid description and 'Taboo word' if a predefined number of players

had the same agreement. The duration of the whole game is 150 seconds and both parties can guess as many images as possible within this time. During the period of four months, the game was played by 13'630 people and 1'271'451 labels for 293'760 images were generated. These numbers show the power of the idea. The players (crowd) did not know what is going on behind the scenes and they also did not realise the purpose of their inputs.

3.8.2 Socialisation

"Social factors such as the desire to feel a sense of involvement and 'belong' to a social group, and the forming and maintaining of interpersonal bounds, are a fundamental human need. Empirical studies also show that social motivation is an important driver for people taking part in online activities, ranging from knowledge contribution to providing emotional support." [26]

One example project which use this social incentive is 'stackoverflow'[5]. People are able to post questions about computer programming issues and other users will provide their help for free. Good answers will receive votes from other contributors and the person who asked the question is authorised to mark an answer as accepted. Hard workers can earn reputation points from other users for questions, answers or edits. A higher reputation score will unlock advanced functionalities. Another way to earn respect from other users is to gather badges. These are achievements which are available in three levels: bronze, silver and gold. "Answer score of 100 and more", "Asked a question with 10'000 views" or "Visited the site each day for 30 consecutive days" are example activities which will be rewarded with badges. The two presented rewards motivate the users of the website to contribute as much content as possible. The community itself is controlling the quality of the answers because experts can remove wrong or low quality statements. Normal users can penalise improper answers by not voting for them. The service sorts answers based on the votes in descending order and the worst evidences will be ignored by the customers.

3.8.3 Unintended by-product

Data from the crowd is collected as an involuntary by-product of the main purpose. One of the most famous projects is reCAPTCHA [23] which is a further development of the well known Captcha[6] idea. The method will show distorted characters, which cannot be recognised by the OCR (Optical Character Recognition) software, to the internet users. The reCAPTCHA acts like a normal Captcha but the inputs will be used additionally to improve text recognition systems.

Another project from the same inventor is Duolingo[7]. Luis von Ahn has the vision to translate every page in the web into every major language. He hides the main purpose of the service behind a free foreign language learning program. Companies remunerate the founder of the project for translated documents.

3.8.4 Financial Reward

Another possibility to attract workers is the good old money. Crowdsourcing platforms offer to pay them for accepted tasks. If the payment is too low, then workers will not process the tasks. High

---

[5]http://stackoverflow.com/
[6]http://www.captcha.net/
[7]https://www.duolingo.com

rewards will attract spammers who deliver bad quality work to collect as much cash as possible in a short amount of time. A research paper from Yahoo [19] investigates the relationship between financial reward, and the performance of the crowd. They found out, that a higher payment increases the quantity of the work and not it is quality. They proposed to use other incentives like enjoyable tasks or social rewards because the quality of work is the same or better than financial driven approaches. A second advice is that requesters should use as less money as possible only if a payment of the workers is possible. Based on the fact that work will be done faster but not better if a higher gratification will be paid.

Amazon itself does not provide numbers but suggests to take a look at similar HITs to compare rewards [2]. A good strategy is also to proof how long it takes to complete the own tasks and then calculate how many tasks can be done in one hour. Different analyses [13,14] show that the median wage is \$1.38/hour and the average wage \$4.8/hour. The Mechanical Turk Tracker website[8] was developed by the author of one of these statistics [14] and it is possible to calculate the average cost per HIT for a specific day. On 10th of March 2014, the website tracked 236'370 completed HITs with a total reward of \$23'110 and an average of \$0.097/HIT. These numbers should help the requesters to find an initial price for their tasks. However, there is no general formula to calculate the right costs for an HIT. If the initial price is too low, the workers will ignore those tasks and try to find others with a better revenue/expense ratio. This results in higher completion times. In this case, the requesters should increase the reward. On the other side if the tasks will be completed very fast and the results are not like expected, then a decrease of the reward can be helpful.

3.9 Demography

The workers on the MTurk platform are hidden behind an identification number, no details about gender or country of residence are available. To get detailed information about the workers, researchers from the University of California published surveys in the form of HITs [21] and presented their results in 2010. They observed the crowd for about 20 months and detected some changes over time. The number of Indian workers raised significantly within one year and approximately one third of the workers were from there. The majority of the turkers was located in the United States (56%) and every tenth in the United Kingdom, Canada or Philippines. The distribution between female and male participants was nearly equal and most of them were between 18 and 35 years old. A very interesting fact is that 41 percent of the workers were highly educated (Bachelor degree). The authors of the paper also provided numbers about the financial situations of the crowd. A fifth needed the money to always or sometimes make basic end meets, 30 percent to buy for nice extras. Unfortunately, the presented facts are four years old but no current numbers are available for the Amazon Mechanical Turk platform.

---

[8]http://mturk-tracker.com

Chapter 4

Implementation

4.1 Technologies

All the produced code is written in Python (Version 2.7.5). Some of the found research papers use the programming language to create for example a workflow with multiple subtasks. There also exists a plugin for the Amazon Mechanical Turk platform called boto (Version 2.25.0). The API has similar functions as the Java SDK provided by Amazon (Listing 4.1). The scikit-learn library (Version 0.13) was used for the machine learning implementation. Therefore, it is possible to create a productive web service with the help of the Django Python web framework in the future.

```python
from boto.mturk.connection import MTurkConnection
from boto.mturk.question import QuestionContent,Question,QuestionForm,Overview,...
from boto.mturk.qualification import LocaleRequirement,Qualifications

title = 'Estimate the price of auction items based on title, description and images'
description = ('Take a look at an item description and estimate the corresponding price')
keywords = 'image, pricing, picture, item, estimation'

mtc = MTurkConnection(aws_access_key_id=ACCESS_ID,
                      aws_secret_access_key=SECRET_KEY,
                      host=HOST)
#-------------- BUILD OVERVIEW -------------------
overview = Overview()
overview.append(FormattedContent(html_code))
#-------------- BUILD QUESTION 1 ------------------
qc1 = QuestionContent()
qc1.append_field('Title','Price estimation (USD)')

qc1.append(FormattedContent(html_code))

fta1 = FreeTextAnswer(None, None, 1)
fta1.constraints.append(NumericConstraint(1, 1000000))
fta1.constraints.append(RegExConstraint("^\+?([1-9]\d*).\d{0,2}$"))

```

```
25  q1 = Question(identifier="price_find",
26                content=qc1,
27                answer_spec=AnswerSpecification(fta1),
28                is_required=True)
29  #-------------- BUILD THE QUESTION FORM -------------------
30  question_form = QuestionForm()
31  question_form.append(overview)
32  question_form.append(q1)
33  #-------------- CREATE THE HIT -------------------
34  qualification = Qualifications()
35  qualification.add(LocaleRequirement('EqualTo','US'))
36
37  hitDetails = mtc.create_hit(questions=question_form,
38                max_assignments=5,
39                title=title,
40                description=description,
41                keywords=keywords,
42                duration = 60*120,
43                reward=0.1,
44                qualifications=qualification,
45                response_groups = ['Minimal'],
46                )
```

Listing 4.1: boto HIT creation example

4.2   Pure Approach

The chapter describes the first of two crowdsourcing approaches. The pure one uses only inputs
of humans.

4.2.1   Ground Truth

Real eBay auctions were collected by the API to generate the ground truth for the crowdsourcing
experiments. The online auction platform divides the items in eight main categories: Motors,
Fashion, Electronics, Collectibles & Arts, Home & Garden, Sporting Goods, Toys & Hobbies, and
Deals & Gifts. The ground truth consists of seven items from every category with the exception
of the Motor's and Deals & Gits sections because the API cannot search for items in these cate-
gories. First, some keywords were created to touch the desired category: "Swiss Watch" (Fashion),
"Smartphone" (Electronics), "Football Trading Card" (Collectibles), "Coffee Machine" (Home),
"Soccer shoes" (Sporting Goods), "Action Figure" (Toys), and "Handbag" (Fashion). The goal
was to have also gender specific and neutral items. An action figure is normally used by male
persons, the handbag by females, and a smartphone by both. The Finding eBay API provides the
method *findCompletedItems* which takes keywords as a parameter and returns a list of completed
auction items. The Python script searches for the first sold item which uses US dollar as currency,
has a description longer than one-hundred characters, and contains at least three images. Only
three images were kept because most of the auctions present the items with a top, front and side
view. Another reason is the clarity for the crowdsourcing tasks. Every ground truth entry has
the attributes title, description, category, condition, price, and image one to three. The table A.1
(page 62) represents the final ground truth for further experiments.

4.2.2 Tasks Workflow

The inputs of the pipeline (Figrue 4.1) are images of the item to sell which were created by the seller. The images from the ground truth are used for the experiments of the thesis. To create item specific information for the auction, four subtasks were designed:

- Generate a title for the auction item.

- Generate a description of the item.

- Find one category for the auction item.

- Estimate the end price for the auction.

The digits in the brackets define the number of assignments. At the end of the pipeline, the sellers receive the information which they need to create an auction on eBay. The starting price of the item depends on the sales strategy of the sellers but the end price should help to find a suitable one. The condition of the item, the auction duration or the payment settings have to be provided by the sellers themselves.
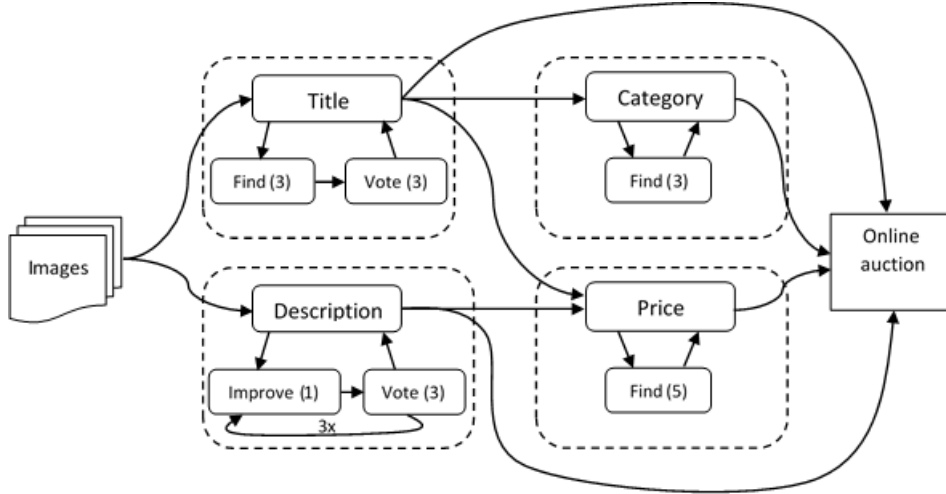
Figure 4.1: Pure crowdsourcing pipeline

4.2.3 Task Design

At the top of every HIT three images of an auction item are shown. Most of the sellers on eBay present their items with a front, side and top view. Only workers from the United States are allowed to participate in the created tasks because the ground truth contains only items from there and they have a better feeling for the currency. Some of the tasks need a voting procedure to determine the final answer. The voters have to mandatorily reason their votes to understand the strength of the selected answer. At the end of every task, the contributors have the possibility to write down a feedback to the requestor.

Workers who did the same HIT in the past are excluded from the one in the future. This should prevent that they can commit the same answer twice. Another restriction is that they cannot vote for the own created solution during the voting task. At the beginning of every task,

a list of worker IDs is shown and a warning that the answers of them will be rejected. This does not avoid the participation of the listed workers by the system but it works because they will not have a lower approval rate. This solution was easier to implement but another one has to be used for a productive solution.

4.2.3.1 Title

The goal of this subtask is to generate a clear and concise title for the auction item with at most eighty characters. The auction platform eBay provides some recommendations for a good title. It should contain the item's brand name, artist, or designer. A specification of the item could also be helpful. The title could include the size, color, condition, and model number, for example. Correct spelling is a must. All these points will be presented to the workers in the instruction section of the task. To make the instructions clear, an example is also listed: After three titles were created

Title: Sony Playstation 4 (black), 1 Controller, New
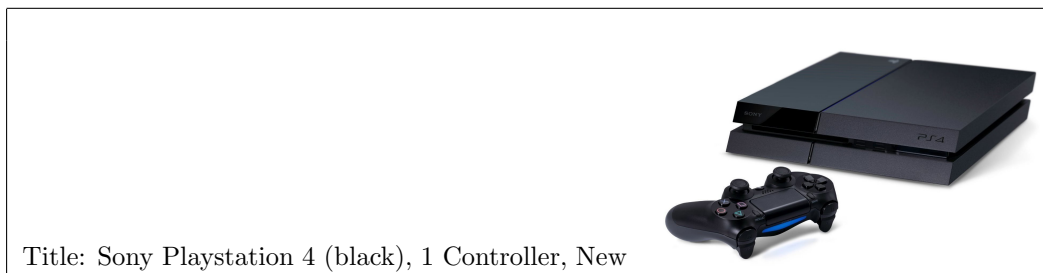
by the crowd, the final title will be elected. If no title receive enough votes, the requester will act as an expert. The expert uses the search engine of the auction platform and decides which title shows more similar items. The turkers will receive $0.05 for finding a title and $0.02 for voting.

4.2.3.2 Description

This subtask is a bit different than the others. An iterative task design is used (Subsection 3.3.2). First, a worker creates an initial description of the item. Then, the second worker can improve the initial solution or create a new one. After that, the crowd decides which description should be kept and which one should be discarded. One iteration includes an improvement and a voting task. Three iterations were used to generate the description of the auction item. The workers should write approximately five sentences and include specific information like size, color, shape, age, manufacture date, company/author/artist, and notable features or markings. TurKit is a Java application to manage iterative approaches. For an improvement of the text a reward of $0.2 will be paid, $0.01 for every participant of the voting procedure.

4.2.3.3 Category

Based on the provided title, the workers have to find the most suitable eBay category for the auction item. The eBay search engine returns one or many categories for a given title. Then, the worker has to decide which one matches best. For making a contribution, the workers achieve a payment of $0.05.

4.2.3.4 Price Estimation

The workers have to guess the end price of the online auction item in US dollars. Reasons for the estimation have to be mentioned additionally. The generated title and description are available for a better understanding of the picture contents. The workers have the possibility to list missing information for a more precise estimation. The participants receive a gratification of $0.05.

4.2.4 Variations

The prior section describes the standard composition of the tasks. To survey the behaviour of the workers, some design modifications were made:

4.2.4.1 Image Quantity and Quality

All available images were presented to the crowd with the highest image resolution. The basis setting of the tasks shows only the first three images. Table 4.1 illustrates the number of additional images per item and the corresponding resolutions.

Ground Truth ID	Total Images	High Resolution (1600 x 1200)
1	6	Yes
2	3	No
3	4	Yes
4	7	No
5	9	Yes
6	3	Yes
7	4	Yes

Table 4.1: Ground truth image quantity/quality

4.2.4.2 Market Price

The actual market price of the items (Table 4.2) was mentioned in the price estimation task. The web service pricegrabber.com was used to find reliable and consistent prices.

Ground Truth ID	Price (in USD)
1	69.00
2	399.99
3	49.99
4	299.00
5	189.99
6	44.99
7	289.99

Table 4.2: Ground truth market price

4.2.4.3 Commission

This section describes the idea of an additional incentive for the workers which is added to the reward of MTurk tasks as a bonus. If an auction item will be sold successfully on eBay, then all

contributors of the crowdsourced result will receive a commission of the end price. The ground truth contains already completed eBay auctions and therefore the criterion of the bonus has to be determined otherwise. Only those created auctions which get more votes during the evaluation process than the ground truth will receive a commission. The table 4.3 shows the distribution of the percentages. The bonus can be between 2.55% and 4.9% of the end price. The range of the end prices in the ground truth goes from $4.99 (watch) to $201 (coffee machine). The commission can be between $0.127 (2.55% of $4.99) and $9.85 (4.9% of $201). The differences of the worker behaviour and a potential quality intensification will be investigated.

Name of task	Number of assignements (Min)	Number of assignements (Max)	Percentage of commission (in %)
Title (Finding)	1	1	0.25
Title (Voting)	2	3	0.1
Description (Improving)	1	1	1.0
Description (Voting)	2	2	0.05
Category	2	3	0.25
Price	1	5	0.5
Total (Min)			2.55
Total (Max)			4.9

Table 4.3: Commission percentages

4.2.4.4 Non-branded Item

All of the ground truth items have visible brand labels. Some are more famous (Apple[1], Puma[2]) than the others (Palisades, Powman Sterling). Information about the brands and their manufactured items can be found easily. Describing an unknown object is more difficult. A non-branded item is put into the pipeline to research the ability of the workers to handle such items (Table A.2, page 62).

4.3 Hybrid Approach

The second approach is illustrated after this introduction sentence.

4.3.1 Ground Truth

A lot of sold items were collected by the help of the eBay API. The used methods were the same as in the prior ground truth generation (Chapter 4.2.1). After all the necessary data was collected, the Python script splits the data shuffled into a training and test set. The training set contains about 70 percent of the whole data. All the consecutively steps (Data analysis, feature ranking) will use the training set until the performance of a classifier will be proved on the test set. The continuous target values are also converted into price classes to use classification algorithms later on. The range of the classes depends on the highest price of the item type. The goal was to

---

[1]http://www.apple.com
[2]http://www.puma.com

generate the same number of classes for every category. The ground truth was generated for three different item types:

Item type	Total number of auctions	Size of training set	Size of test set	Price class range (USD)
Apple iPhone	2'299	1'609	690	25
Hot Wheels Cars, 1:64, Ford Mustang	945	661	284	2
Sony Playstation	943	660	283	25

Table 4.4: Ground truth sets for machine learning

4.3.2 Tasks Workflow

The hybrid pipeline (Figure 4.2) works the same as the pure one except that two subtasks are supported by machines.
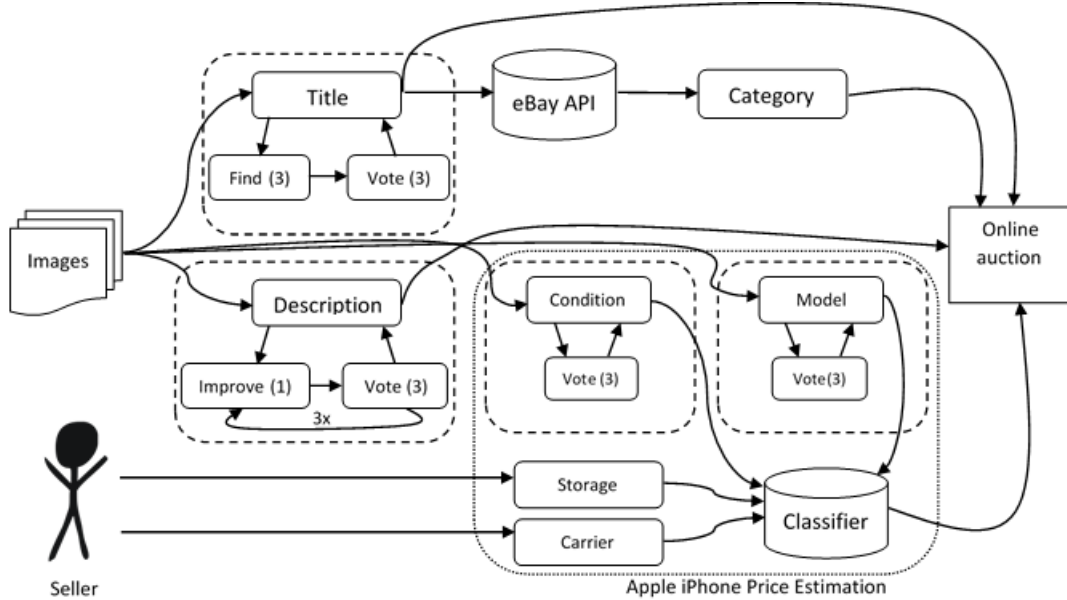
Figure 4.2: Hybrid crowdsourcing pipeline

4.3.3 Task Design

The design of the tasks is similar to the pure approach except of two subtasks: Category and price.

4.3.3.1 Category

The implementation finds the most suitable category by using the eBay Finding API based on the output of the title task. Most of the titles are too specific and the API does not return a category. The algorithm reduces the number of words until a category is found. The used method returns a sorted histogram of categories.

4.3.3.2   Price Estimation

The goal of the task is to estimate the end price of the auction by using a machine learning approach. The input features are built by a combination of crowdsourced and user specified information. The pipeline shows the required fields to estimate the price of an iPhone. The model and the condition of the item is determined by the crowd. The seller provides the information which is not possible to collect by others. For example, the storage size of a phone which is not visible. If all needed features are available, then the machine learning algorithm will produce an end price.

4.3.4   Pre-processing

During the collection of the sold items, some pre-processing steps were necessary to produce accurate results by the machine learning algorithms. All the recorded features were normalised within the parameters of 0 and 1 to generate a uniform feature space. The target labels (price) remain unaffected. Another problem was the quantity of items for a single auction. The quantity field of the entry was one, but the auction contains a lot or a set of items. If an auction title contains a certain keyword ("Set", "Lot", "Pack" or "Bundle"), then the entry will be ignored. Another problem are the inconsistencies of the item descriptions. For example, some auctions contain different declarations of the model in title or description. These inputs were ignored too.

4.3.5   Feature Extraction

The extracted features are divided into three categories.

4.3.5.1   Item Specific Features

The features of this subsection are dependent of the present item category. The number of features is reliant on the available item specific fields provided by the eBay system.

Apple iPhone   The iPhone made by Apple is available in eight models. The first generation was released in 2007, the last model 5S in 2013. Every model comes with different storage sizes (from 8GB to 64GB). The values for the condition property on eBay depend on the corresponding item category. All the values are nominal and will be converted to numerical.

Mattel Hot Wheels Cars   Mattel[3] produces diecast car models in different sizes and series. Cars with a ratio of 1:64 are the most popular ones. The collected data contains only Ford Mustang cars because they are very famous in the US and it should be possible to distinguish between different models. The exact model is indicated by a date.

Sony Playstation   Sony's[4] Playstation exists in ten versions. Two of them are portable and for the second and third model of the console, a slim version is available. Every device has a region code or all data carriers are readable. There is no simple way to extract the storage size of the consoles at the moment because eBay does not provide a field for this specific information.

---

[3]http://www.mattel.com
[4]http://www.sony.com

Name	Description	Values	Range	Data type
Model	The model of the iPhone where 0 is the oldest generation and 7 the newest	1st, 3G, 3GS, 4, 4S, 5, 5C, 5S	[0, 7]	Integer
Storage	The size of the storage of the smartphone	8GB, 16GB, 32GB, 64GB	[1, 4]	Integer
Condition	The condition of the iPhone	New, New other, Manufacturer refurbished, Seller refurbished, Used, For parts or not working	[1, 6]	Integer

Table 4.5: iPhone specific features

Name	Description	Values	Range	Data type
Model	The model of the Ford Mustang where 1964 is the oldest and 2014 the newest		[1964, 2014]	Integer
Condition	The condition of the car	New, Used	{1, 2}	Integer

Table 4.6: Hot Wheels specific features

Name	Description	Values	Range	Data type
Model	The model of the Playstation	1, 2, 2 Slim, 3, 3 Slim, 4, Vita, Portable	[0, 7]	Integer
Region Code	The region code of the console	Not specified, NTSC, PAL, Region free	[0, 3]	Integer
Condition	The condition of the item	New, New other, Manufacturer refurbished, Seller refurbished, Used, For parts or not working	[1, 6]	Integer

Table 4.7: Playstation specific features

4.3.5.2 Auction Specific Features

The auction itself is described by the features in this section. The list (Table 4.8) contains some timing and shipping information. The number of pictures and the description length could also have an influence on the result of the auction. All values are numerical.

Name	Description	Range	Data type
Duration	The duration of the auction in days	{1, 2, 3, 7, 10}	Integer
Number of pictures	Number of pictures attached to the auction	[1, 12]	Integer
Length of description	Length of the item description	[0, 500'000]	Integer
End weekday	The last weekday of the auction duration	[1, 7]	Integer
Start weekday	The weekday of the creation date	[1, 7]	Integer
End hour	At what hour the auction was ended	[0, 23]	Integer
Global shipping	The item will be shipped over the whole world or not	{0, 1}	Boolean
Shipping locations	The number of countries where the item will be shipped	[0, 249]	Integer
Shipping type	Specifies the calculation of the shipping costs	[0, 7]	Integer
Returns accepted	If the buyer can return the item or not	{0, 1}	Boolean
Handling time	How many days it will take until the item is put in the mail once the seller receive payment	{1, 2, 3, 4, 5, 10, 15, 20}	Integer

Table 4.8: Auction specific features

4.3.5.3   Seller Specific Features

These features (Table 4.9) characterise the seller who created the auction. Every user on eBay has the possibility to give a positive, neutral or negative feedback after every transaction. The rating system awards stars with twelve different colors for trustful sellers. After ten positive feedbacks, the user receives a yellow star for example. Therefore, the nominal value has to be converted to an integer.

4.3.6   Data Analysis

This chapter takes a closer look at the collected data. Every item category has an attribute model. The price of the smartphones depends on the date of appearance (Figure 4.3). The newest iPhone

Name	Description	Range	Data type
Seller rating	Percentage of positive feedbacks	[0, 100]	Float
Seller rating count	Number of positive minus negative buyer feedbacks	[0, 12]	Integer

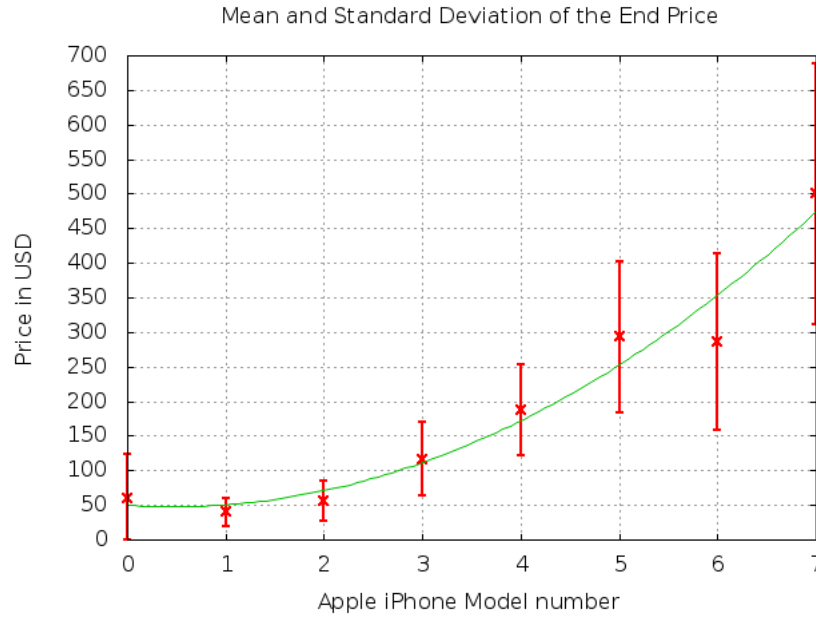Table 4.9: Seller specific features

Figure 4.3: Model/Price scatter plot (iPhone)

5S produces the highest price, the second generation the lowest one. The debut feature produced by Apple has a high value for collectors and are traded higher than some later versions. The mean values of every iPhone model can be roughly estimated by a quadratic function. The data of the Playstation shows similar characteristics for the average prices related to the price feature. The Hot Wheels Cars have a completely different price distribution. There is no visible pattern for the recorded models. The Ford Mustang from 1983 has the highest average price of $27.02 but was sold only once. The rarity of the items has a higher influence on this category than to the other ones.

The histogram of the iPhone (Figure 4.4) illustrates that the model 4 and 4S are involved in about 60% of the collected auctions. The Playstation 3 Slim is the most dominant console (about 39%). The previous model of the actual one is dominating the online marketplaces. The 1967 and 1971 are the most popular Mustang models in the data set with 31 different types.

An executed feature ranking [11] with SVMs indicated surprisingly that some features are not as important as expected. The model of a car has only a small influence on the end price. The condition of the item and the duration of the auction are important for all the three tested categories.
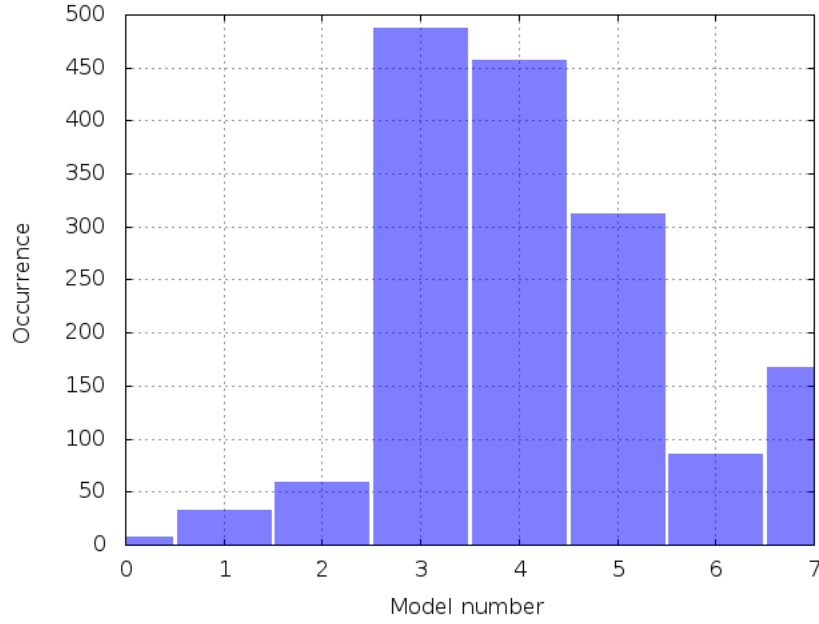
Figure 4.4: Model histogram (iPhone)

4.3.7   Machine Learning Algorithms

Three machine learning algorithms were used to estimate the price. The theoretic knowledge about
the algorithms is given in the following of this preface.

4.3.7.1   k-Nearest Neighbours

The kNN algorithm [6] represents every sample of the training set in an n-dimensional feature
space where n are the total number of features. The class correspondence of the data points
is stored too. For the classification of a test sample, the k-nearest neighbour data points are
determined. Usually, the Euclidian distance is used to calculate the distance between the points
in the n-dimensional space. The data point is assigned to the class with the majority in the
neighbourhood. If no class is dominant, then the k is decreased by one until the tie is broken. The
standard configuration of the algorithm uses uniform weights for the data points. This means that
each point in the neighbourhood has the same influence on the result. Another way to determine
the weight of a neighbour is to calculate the inverse of the distance to the point under supervision.
The algorithm can also be used for continuous values (Regression). In that case, the average value
of the k-neighbours will form the regression output.

Sellers of online auction items compare previous auctions for the same or similar ones to estimate
the price. The kNN algorithm works similar and should provide good results.

4.3.7.2   Multiclass Support Vector Machines

Normally, Support Vector Machines (SVMs) [5] are used for binary decisions. To classify multiple
classes, the "one-versus-one" approach is used. If there exist three classes in total, three SVMs
(Class A vs. class B, class A vs. class C, class B vs. class C) are needed. The class with the
majority of the votes will be the resulting output. The idea of the classifier is to map the inputs

into a high-dimensional feature space for an accurate separation by one or more hyperplanes. The hyperplanes can be linear or non-linear (e.g. Polynomial, Gaussian). The margin between the two classes should be maximised. Points on the margin are called support vectors. These vectors have a higher influence on the classification. The class membership of an input sample is determined by the location (relating to the margin) of the point in the high-dimensional feature space. A regression based implementation of the algorithm is available as well [8].

Some online auctions for the same item produce a higher price of sale than others (outliers). The goal is that such observations have no or only a small influence on the price prediction. SVMs use a subset of points (Support vectors) to determine the class membership. Other points which are far away from the margin have no influence on the final result and the auction outliers should play this role. Therefore, the SVMs could be a good solution for the discussed problem.

4.3.7.3   Random Forest Classifier

The Random Forest classifier was introduced by Leo Breiman in 2001 [4]. The classifier combines multiple randomised decision trees and averages their results for a final decision. The size of the forest is one of the parameters of the classifier. The number of features considered for a split node another one. The calculations of the outputs can be parallelised because all the trees are considered. The paper explains the creation of the trees, the training procedure and gives the mathematical background to understand all the information.

The price of an item is mostly dependent on the number of features and the quality or quantity of them. If a certain feature is available, then the seller expect that the end price of the auction will be higher than without this feature. A car with an integrated air conditioning will be sold for a higher price than the same one without an air conditioning, for example. Therefore, a decision tree should help to create a decision process based on the available features which seems like a natural human behaviour. One tree alone is not enough to cover all the different circumstances.

4.3.8   Parameter Search

A grid parameter search was done to find the best parameters of the introduced machine learning algorithms. The idea of this approach is to train a given classifier to predefined sets of values and keep the ones with the best performance. A 5-fold cross-validation was used to generate separate training and test sets. The original test set stays untouched. The method splits the set into five equally sized subsets. Four subsets are used for training and one for testing, then the roles change clockwise until every subsets was used as test set. The final result is calculated by the average performance of every iteration. The procedure helps to avoid overfitting of the models. If an algorithm is more accurate in fitting known data but less accurate in predicting new data, then the model is overfitted.

4.3.9   Signifigance Tests

The significance test should help to find out if the results of two classifiers are happened by chance. First, the null hypothesis H_0 has to be formulated:

"The mean performance of classifier A is the same as classifier B"

The hypothesis can be rejected if the calculated p-Value is lower than 5%. The value represents the probability that the null hypothesis is correct. The results of the classifiers are not normally distributed, therefore the following tests were used.

4.3.9.1   G-Test

The G-Test [12] is used for the nominal labels and is appropriate for multiple classes. It is a modification of the Chi-Squared test but can handle smaller observed frequencies in a cell of the contingency table. Not every price class occurs in the outputs of the classifiers, therefore the G-Test is favoured over the Chi-Squared test. The results of two classifiers are grouped into a 2 x N contingency table where the rows represent classifiers A and B. N is the total number of different classes in both results. The outputs of the algorithms A and B are recorded at every cell in the table. After that, the expected frequency is calculated for every cell. Based on these two tables the G-Test algorithm calculates the corresponding p-Value.

4.3.9.2   Wilcoxon-Signed-Rank Test

The Wilcoxon-Signed-Rank test [24] is an alternative to the paired t-test but assumes that the population is not normally distributed. The test verifies, if the difference between the two given outputs of the regression algorithms is symmetric about zero. First, the absolute differences will be sorted in ascending order. Then, the samples receive a rank starting with the smallest as 1. After that, a p-value will be calculated.

Chapter 5

Evaluation

5.1 Pure Approach

The section summarise the results of the pure crowdsourcing experiments. First, the overall performance of the pipeline is evaluated. After that, the results of the four subtask are assessed.

5.1.1 Overall Performance

After all the necessary content was created by the crowd, the evaluation was done by comparing the ground truth with the generated information. The evaluation task presented the field title, description and category to the workers. They had to decide which auction description is the best one in their opinion. The result is illustrated in figure 5.1. The commission approach got a majority of the votes with 57.14%. The generic item received all five votes from the crowd. The ground truth contains unnecessary information and looks like a scam, are the statements of the participants.

The workers had the possibility to reason their decision. They motivate their votes as follows:

- Higher value of information.

- Professionalism.

- Hidden information is given (e.g. size of the cleats, size of the smartphone storage).

- The description is clear, short and to the point.

- Authenticity of the article.
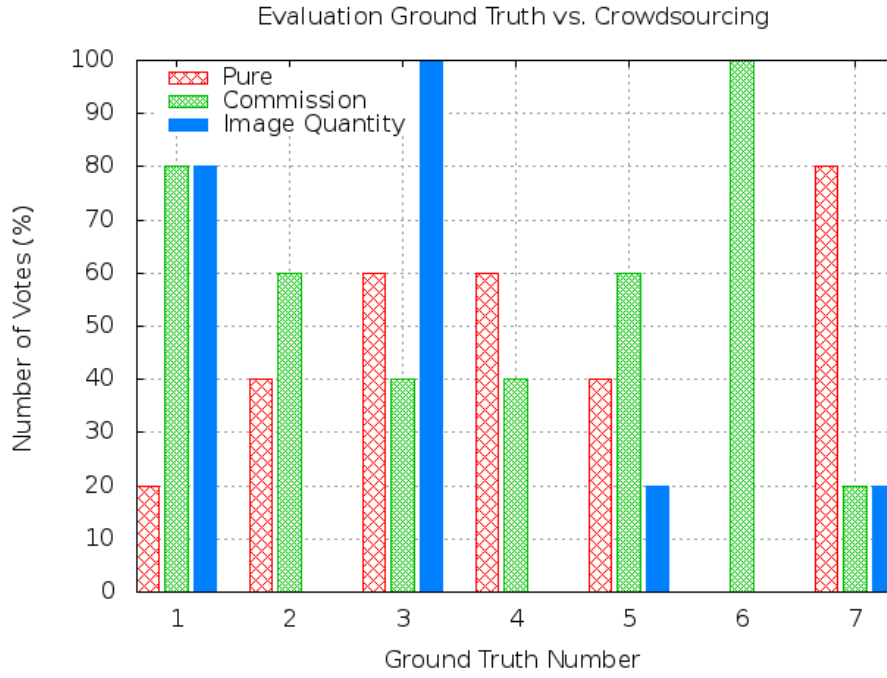
- Grammatical issues.

Figure 5.1: Evaluation of ground truth vs. crowdsourcing

5.1.2 Title

The length of the title (Figure 5.2) increases if the task contains all available pictures of the item. This setup produced an average title length of 64 characters. The other experiments have an average of 50 for the standard design and 48 for a promised commission. The median value shows almost the same ranking except that the additional reward in form of a commission generates more characters than the standard configuration. The item without a brand was described by a title with 37 characters. The length does not say anything about the quality of the title but it indicates the influence of the different task settings. The usage of natural language processing tools could help to get more information about the quality but this would be outside the scope of the thesis. The workers spent more time to find a title if they can achieve an additional bonus (Figure 5.3). The average working time (203.76 seconds) was twice as much as without a commission (101.76 seconds). If the writer of the title had to look at more than the three standard images, then they needed more time to commit a title (131.28 seconds). If the additional time effort was used to open all the supplementary images is not clear at the moment.

After the creation of the titles, the workers had to select the final one for each item. They justify their selections as follows:

- Amount of details.

- Attracts more attention.

- Research on eBay produces better results for the title.

- Experience with online auctions.
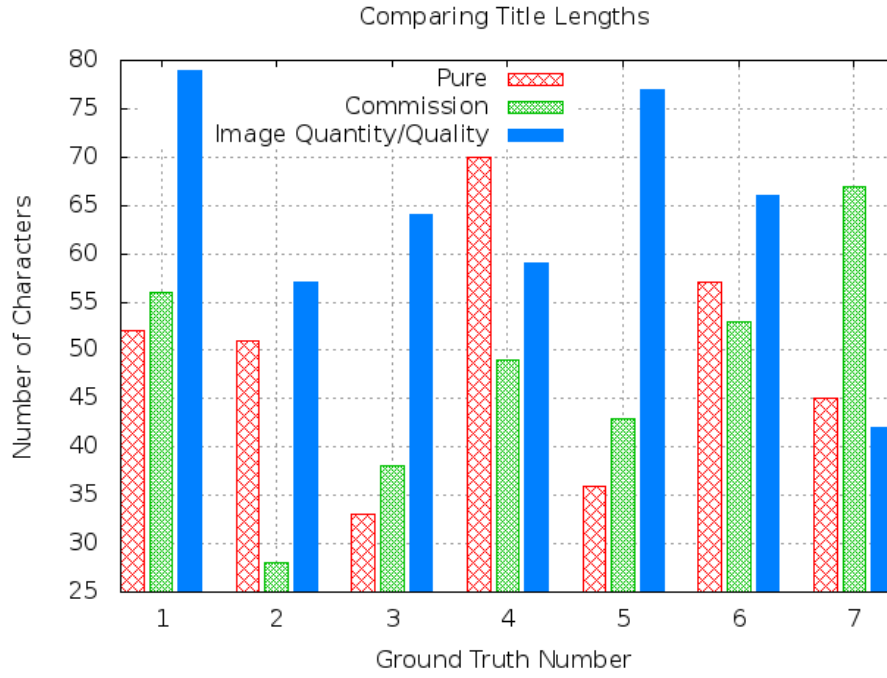
- Wrong information.

Figure 5.2: Evaluation of title lengths

- Too long.

5.1.3 Description

The same measurements as for the title were done for the description (Figure 5.4). The general task design achieved the highest average length (467 characters) but it contains an outlier with 1'706 characters for the ground truth item number 5. The worker copied an item description from the website of the producer. More images does not conclude to a longer description because the average length of these titles is 281. An extra reward leads to 455 characters, but has the highest median of 572. The numeric parameter is more robust to outliers. The other setups follow by 307 (Commission) and 238 (Pure). The non-branded item realises a length of 240 characters which is equal to the median of the branded items. Therefore, the workers are able to write descriptions of equal length regardless of which item type (branded/non-branded) was presented to them.

5.1.4 Category

Most of the workers submitted only the main category and not the complete hierarchy with main and sub categories. For the football trading cards they wrote "Sports Mem, Cards & Fan Shop" instead of "Sports Mem, Cards & Fan Shop:Cards:Football", for example. One reason could be that the instructions were not clear and precise enough. The desired format of the input was not mentioned too. An improved task design will result in more accurate outputs.
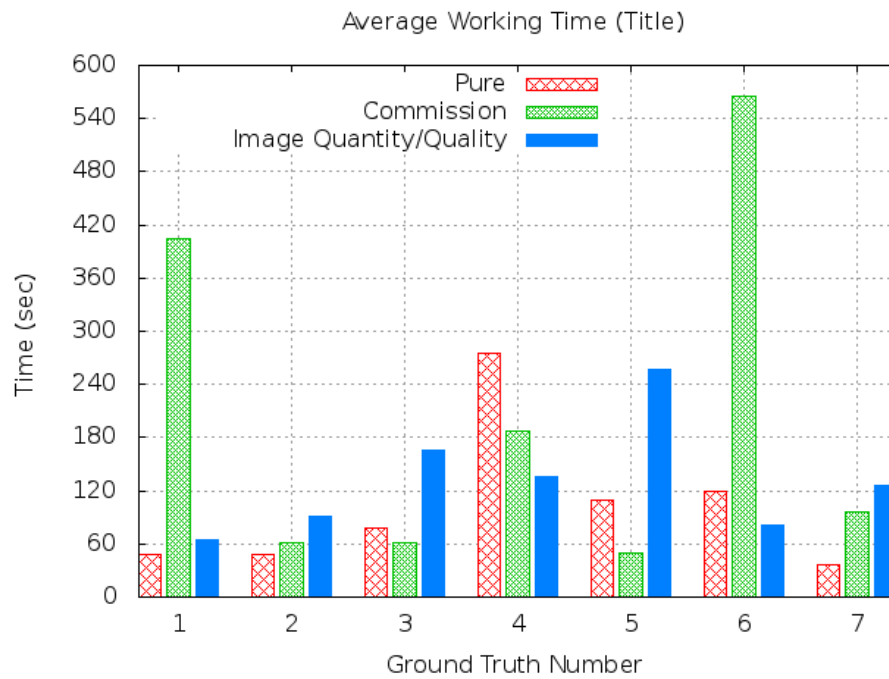
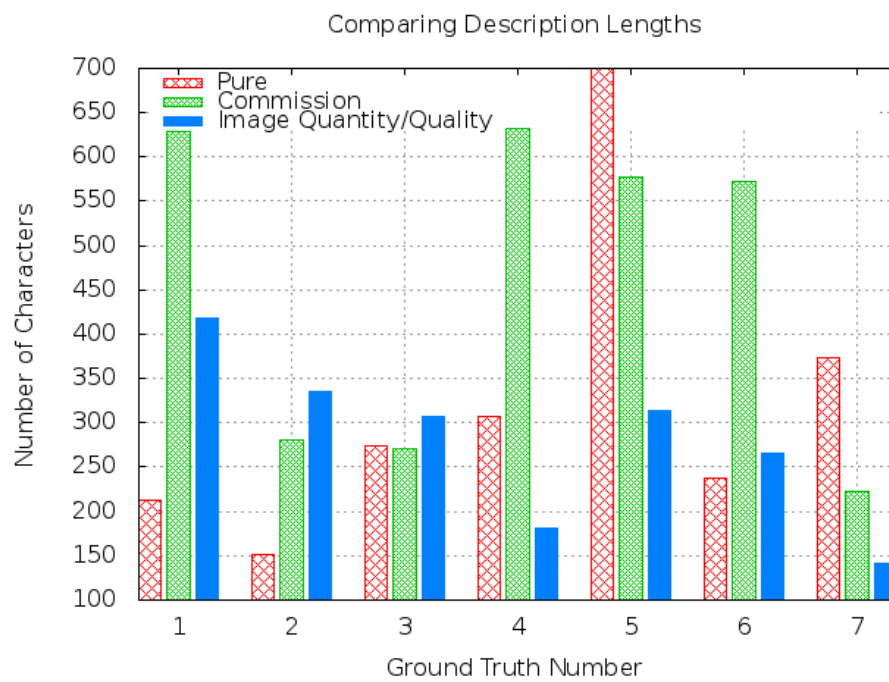Figure 5.3: Evaluation of the average working times for finding a title

Figure 5.4: Evaluation of description lengths

5.1.5 Price Estimation

To evaluate the accuracy of the price estimations (Figure 5.5), the root mean squared error (RMSE) was used. It is frequently used to measure the quality of predictions. The true values are the end prices from the ground truth. If the workers have the actual market price of the item at one's disposal, then they predict the price best with an average RMSE of 51.46 USD. The experiment with the commission ranked just behind with 58.55 USD. The others appear with 81.75 (Pure) and 89.89 (Image quantity/quality) at the end of the ranking. The tested item without a brand has a RMSE of 1074.12 USD. The comparison between branded and non-branded items can be done by dint of the absolute percentage error. The value represents the difference between predicted and exact value as a percentage of the exact value. The seven ground truth items have an error of 51% in median, the non-branded item one of 83%. The values from the branded items are taken from the standard experiment to have the same preconditions for both categories. The results indicate the challenge of the price prediction for generic things.

The digital watch and the handbag have a lot of predictions which are higher than the ground truth price (Figure 5.6). Reasons could be that the watch was sold at a cheaper price than expected and the handbag could be a fake.

The estimation errors can be explained by different observations. The size of the smartphone storage is not given. The price difference of an iPhone 5S with 32GB and 64GB is about 100 dollars. The football trading cards are difficult to estimate because most of the workers do not know the football players and cannot asses the rarity or popularity of them. The football boot model is available in different versions. The original one and cheaper replicas. The action figure is rare and no other auctions are available to compare with.

The results of the expenditure of time for the price estimations (Figure 5.7) are disappointing. The workers spent between 155 (Pure) to 493 (Image quantity/quality) seconds on average to find an appropriate price for the items. The promised commission had only a marginal influence on the time statistic.

5.1.6 Variations

5.1.6.1 Commission

The ground truth items 1 and 3 received a majority of the votes from the crowd. The commissions were paid manually using the web interface of the Amazon Mechanical Turk web service. The value of the bonus has to be shortened to two digits after the point and rounded up to $0.01 because of some restrictions of MTurk. The bonus aggregates to $1.48 for both auctions (see table B.1, page 63). The workers also received an additional message:

"Dear worker,

You receive a commission (0.25% of the end price) as bonus payment for your work. The end price of the eBay online auction was $27."
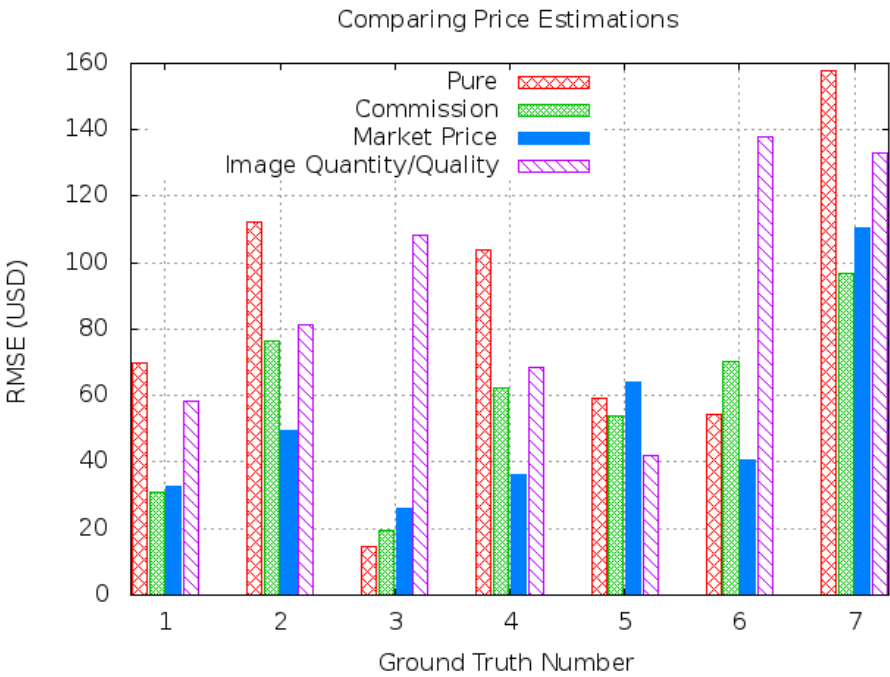
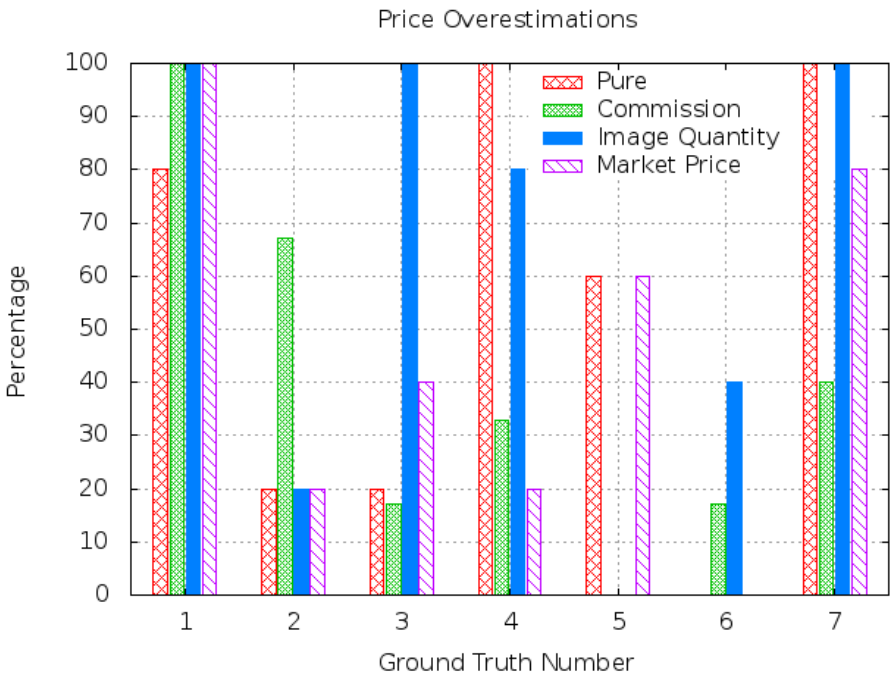Figure 5.5: Price prediction quality
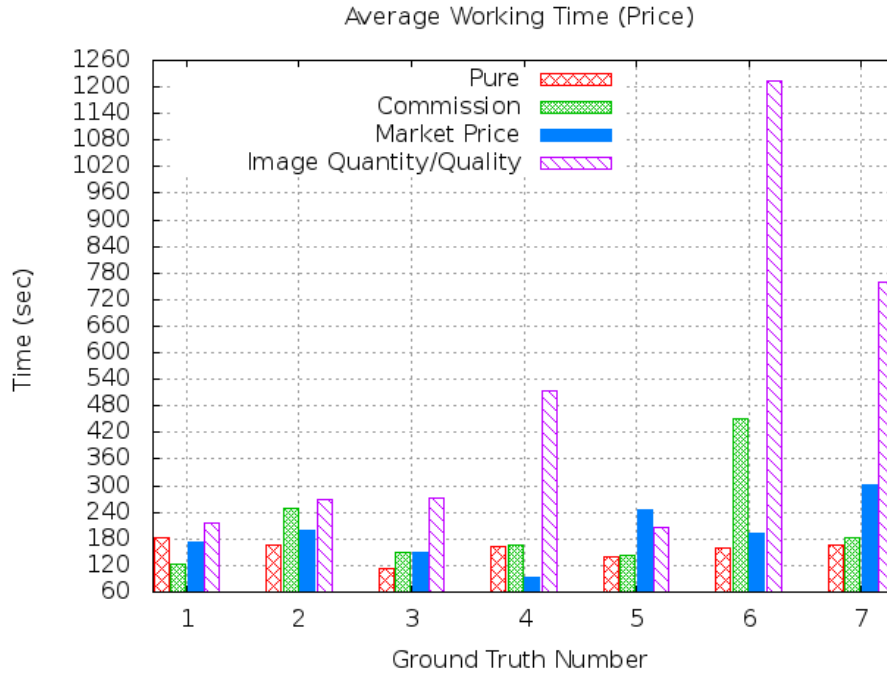

Figure 5.6: Price overestimations

Figure 5.7: Evaluation of the average working times for the price estimation

5.2 Hybrid Approach

The Random Forest approach shows the best performance over all three item categories independent of classification or regression (Figure 5.10/5.12). The results of three iterations were averaged because of the randomness of the classifier. The accuracy of the classifier depends also on the item category. The RFC assigns every third input to the right Sony Playstation price class. The plot of the mean absolute error (MAE, Figure 5.11) assures the strength of the classifier and visualises how far away the predictions are on average. kNN has an error of about four classes for the iPhone category. One class has a range of 25 USD.

The reasons for the moderate performance of the classifiers could be multi-object and/or outdated auctions in the collected data. A higher number of features could also lead to better results.

The comparison between the human-based and the machine-based prediction is not possible with the aid of the RMSE because the values are scale-dependent. The random forest algorithm predicts a final price of the ground truth item number 2 (Apple iPhone) of $108 and the crowd a price of $172. The crowdsourced price is taken from the experiment where the workers had access to the market price because this experiment has the lowest error rate. The ground truth price is $185. The crowd has an absolute price difference of $13, the machine learning approach one of $77. The mean absolute error confirms this observation. Five workers predicted prices with an error of 21%, the machine learning approach one of 30.9%.

The normalised confusion matrix (Figure 5.9) of the iPhone classifiers illustrates the distribution of the assignments. The origin is located at the left top position. The x-axis represents the truth-values and the y-axis the values assigned by the classifier. The values were normalised per row. A perfect classifier would produce a diagonal which contains only ones. The scatter plot (Figure 5.8) is another way to visualise the relations between the truth and predicted values. The criteria of a
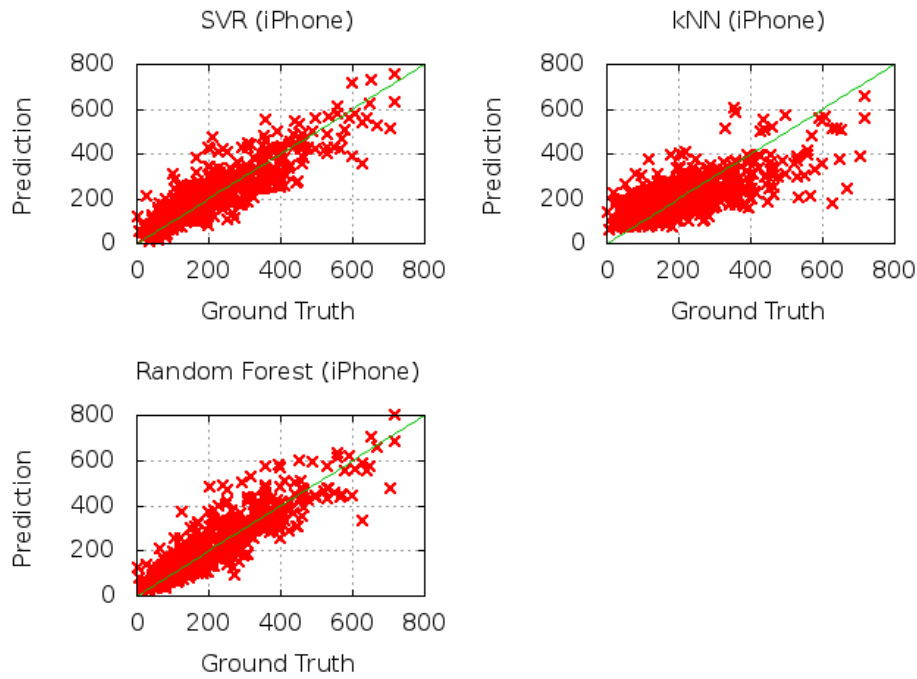
Figure 5.8: Scatter plot true vs. prediction (iPhone)

perfect classifier are the same as for the confusion matrix.

The correlation between the two values can be measured by the Spearman rank-order coefficient (Table 5.1). The measured coefficient can take values from -1 to +1. 0 implies no correlation, a positive correlation implies that as the true values increase, so do the predicted values. If one variable decreases and the other one increases, then the correlation value will be negative. The results confirm the power of the random forest algorithm.

The plots for the other categories and the exact accuracies/RMSEs/MAEs are enumerated in the appendix section.

Algorithm	iPhone	Mustang	Playstation
kNN	+0.689	+0.138	+0.589
SVR	+0.859	+0.191	+0.742
Random Forest	+0.875	+0.521	+0.879

Table 5.1: Spearman rank-order correlation coefficients
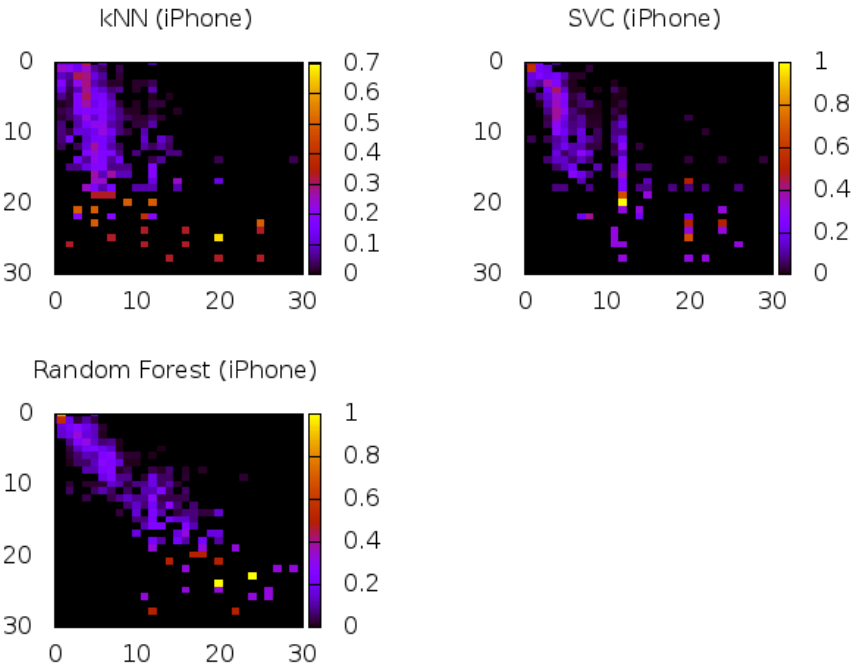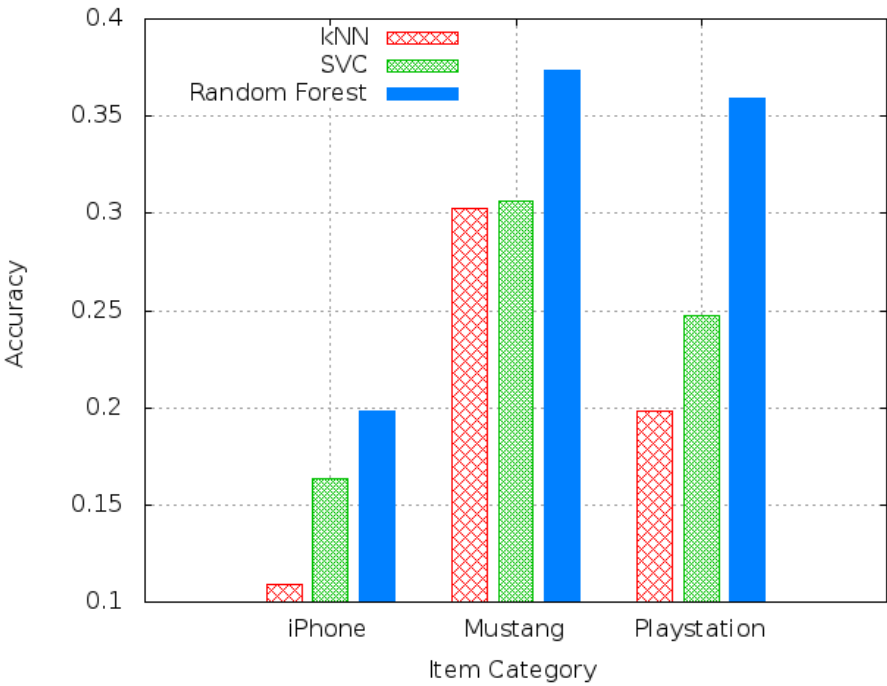
Figure 5.9: Normalised confusion matrix (iPhone)
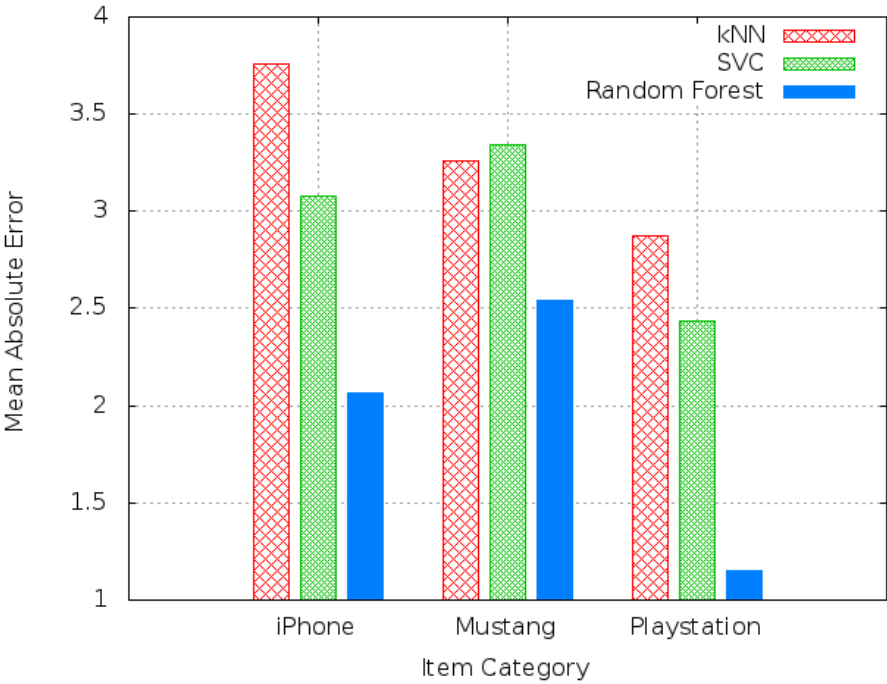

Figure 5.10: Classification accuracy

Figure 5.11: Classification mean absolute error

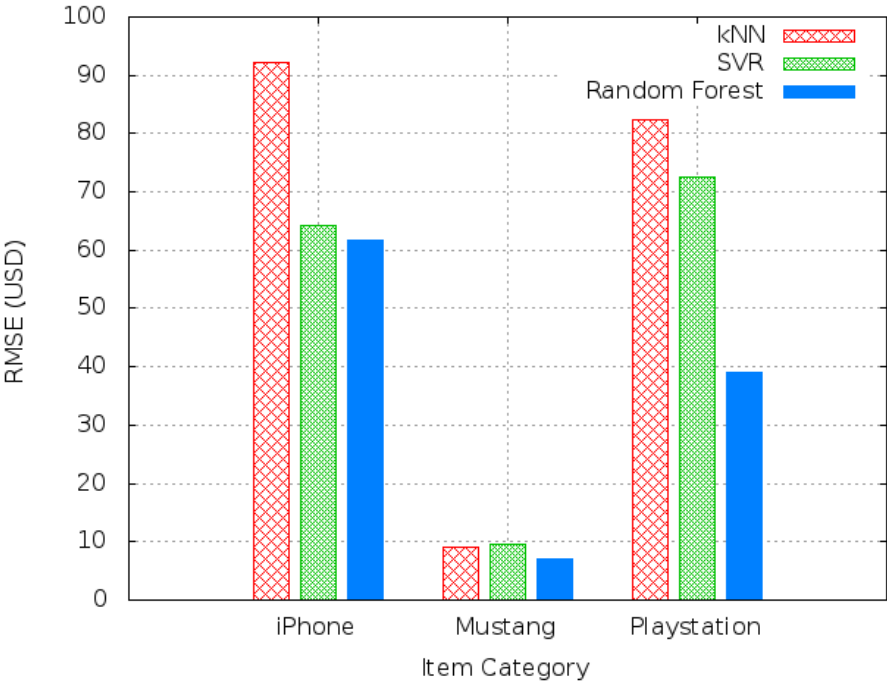Figure 5.12: Regression root mean squared error

Chapter 6

Discussion

The hope of an additional reward in form of a commission leads to the best overall result. A majority of the crowd favours the created contents over the ground truth. This setting guides to longer item descriptions and a more precise end price prediction.

A supplementary set of images has no influence on the quality of the auction descriptions. This scenario directs only to longer titles and descriptions. The workers spent more time to answer the questions on average but this additional effort is not enough to affect the quality of the results.

The price estimations are more accurate if the workers have access to the market price of the items. This experiment has the lowest root mean squared error. A combination of a promised bonus and an available sales price could drive to the best result in the future.

A non-branded item results in a shorter title and a higher price prediction error. However, the workers used twice as much time as for branded items.

The crowd predicts the end price of an Applie iPhone more precisely than the implemented machine learning approaches. They can find current auctions on eBay to compare with and can detect actual trends. The machine learning algorithms have to be retrained from time to time because eBay is a real marketplace and the prices will develop over time. Another drawback of the hybrid idea is the specialisation in specific item categories, a certain smartphone model for example. The price advantage of the implementation is negligible in comparison with the accuracy of the crowd.

The human-based approach finds the most suitable eBay category for the items, but the workers were not able to transfer the correct name to MTurk (Spelling errors). The hybrid solution can be done for free but the resulting category is not as precise as the human-powered implementation.

Chapter 7

Conclusion

7.1 Future work

7.1.1 Google Reverse Image Search

Some tests during the first phase of the thesis has shown that Google's reverse image search does not return reliable results for all types of items. If the search algorithm find websites which contain information about the item, then the application should extract the relevant facts in a certain way. Some experiments should be done and the possibilities of the API should be investigated in the future.

7.1.2 Main Image Selection

The workers have to decide which of the available images give the best résumé of the product. If the application should publish the auction on eBay in the future, then a representing image of the item has to be determined.

7.1.3 Fully Automated Application

The creation, administration and evaluation of every subtask is done manually at the moment. The final product can be a mobile application and/or a web service which manages the whole process. The user will take some pictures of the item and upload the data to the server. Then, he has to provide some auction specific information (duration, shipping details) and the software will create the auction after all missing inputs are generated by the crowd. Different pricing strategies can be selected. The estimated price can be used as the starting price or the price can be reduced by a predefined percentage rate.

7.1.4 Price Estimation Game

Another idea to estimate the starting price is inspired by a German TV game show. The candidate has to predict the cost of an article. After the first guess, the game master answers with 'higher' or

'lower' until the right guess occurs or the time is running out. If the player finds the correct price, then she/he will win the object. The idea of the show is modified to implement a game with a purpose similar to the ESP game project [22]. The general procedure of the game is the following:

1. The system waits until two independent players are connected and ready to play.

2. A few pictures, title, and description of the article are displayed and the players had to read them first.

3. Then, the game starts and a first guess of the price will be shown by the system.

4. Both users have to decide if the real price is higher or lower than the displayed one.

5. Dependent on the previous response, the system will present a higher or lower price until the countdown is expired or there are no guesses left.

6. The players will receive a score dependent on the difference of the price estimation. A smaller difference leads to a higher score, a higher one to a lower score.

The first guess of the system will be the mean value μ of a large number of sold items on eBay. The value can be determined by the eBay API. The guessing structure will be implemented as a directed binary tree. The root node represents the mean value and every following child node will have a lower (left child) v_l or higher (right child) value v_r determined by the value of the parent node v_p and the depth d of the tree. The following formula calculates the values of the nodes:

$$v_l(v_p, d) = v_p - \frac{\mu}{2^d} \tag{7.1}$$

$$v_r(v_p, d) = v_p + \frac{\mu}{2^d} \tag{7.2}$$

The leafs are integer values which cannot be divided by two and represents the final guess of a player. If the time is up and the guesser does not reach a leaf node, the value of the actual node is taken. The score of the price prediction is determined by a scoring function s where x_1 and x_2 are the price estimations of player 1 and 2.

$$s(x_1, x_2) = 1 - |\varphi(x_1) - \varphi(x_2)| \tag{7.3}$$

The function φ is responsible to normalise the estimations (interval from 0 to 1).

$$\varphi(x) = \frac{x}{2\mu} \tag{7.4}$$

The function is also used to weight the different estimations for the same product. If n rounds were played for a given object, the final price t will be calculated:

$$t = \frac{1}{\sum_{k=1}^{n} s(x_{k1}, x_{k2})} \left(\sum_{i=1}^{n} s(x_{i1}, x_{i2}) \frac{x_{i1} + x_{i2}}{2} \right) \tag{7.5}$$

The reliability r of the price estimation is the mean score of all played games for the same object:

$$r = \frac{1}{n} \left(\sum_{i=1}^{n} s(x_{i1}, x_{i2}) \right) \tag{7.6}$$

The formulas of the presented idea are the results of a brainstorming and have to be proven first.

7.2   Pros and Cons

This section states the assets and drawbacks of the thesis idea from the viewpoint of the author.

7.2.1   Pros

Workers are impartial and enumerate the facts of the item based on the pictures. They do not try to write a sales text. The provided photos are not ideal to describe the peculiarities of the item. The worker has no chance to mention the storage size of a smartphone if no screenshot of the system properties is given. If the photographer follows some guidelines to catch the properties of the item, then the workers can deliver good work. The participants of the price task have fun to guess the most accurate price. This fact has a positive influence on the prediction results.

7.2.2   Cons

Workers are lazy and minimalist. They copy item descriptions from the websites of the producers and take the first found price of sale for the estimation. The prior goal is to maximise the hourly wage and not to commit high quality content. Only the owner of the items knows about small defects and the peculiarities of the item. This important information is missing in the final description and does not allow to create an accurate public sale. The spelling of the written descriptions is improvable.

Bibliography

[1] Omar Alonso and Matthew Lease. *Crowdsourcing for Information Retrieval: Principles, Methods, and Applications.* In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1299–1300. ACM, New York, NY, USA (2011). ISBN 978-1-4503-0757-4. doi:10.1145/2009916.2010170. URL `http://doi.acm.org/10.1145/2009916.2010170`.

[2] Amazon. *Requester Best Practices Guide.* URL `http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf`.

[3] Michael S. Bernstein/ Greg Little/ Robert C. Miller/ Björn Hartmann/ Mark S. Ackerman/ David R. Karger/ David Crowell/ and Katrina Panovich. *Soylent: A Word Processor with a Crowd Inside.* In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 313–322. ACM, New York, NY, USA (2010). ISBN 978-1-4503-0271-5. doi:10.1145/1866029.1866078. URL `http://doi.acm.org/10.1145/1866029.1866078`.

[4] Leo Breiman. *Random Forests. Mach. Learn.*, 45(1):5–32 (October 2001). ISSN 0885-6125. doi:10.1023/A:1010933404324. URL `http://dx.doi.org/10.1023/A:1010933404324`.

[5] Corinna Cortes and Vladimir Vapnik. *Support-Vector Networks. Mach. Learn.*, 20(3):273–297 (September 1995). ISSN 0885-6125. doi:10.1023/A:1022627411411. URL `http://dx.doi.org/10.1023/A:1022627411411`.

[6] T. Cover and P. Hart. *Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theor.*, 13(1):21–27 (September 2006). ISSN 0018-9448. doi:10.1109/TIT.1967.1053964. URL `http://dx.doi.org/10.1109/TIT.1967.1053964`.

[7] CrowdFlower. *Case study - eBay* (2013). URL `http://cdn2.hubspot.net/hub/346378/file-522132326-pdf/docs/CF-eBay-CS.pdf?t=1392311997000`.

[8] Harris Drucker/ Chris/ Burges* L. Kaufman/ Alex Smola/ and Vladimir Vapnik. *Support vector regression machines.* In *Advances in Neural Information Processing Systems 9*, volume 9, pages 155–161 (1997). URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.5909`.

[9] Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara. *Towards an Integrated Crowdsourcing Definition. J. Inf. Sci.*, 38(2):189–200 (April 2012). ISSN 0165-5515. doi:10.1177/0165551512437638. URL `http://dx.doi.org/10.1177/0165551512437638`.

[10] Rayid Ghani. *Price Prediction and Insurance for Online Auctions.* In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 411–418. ACM, New York, NY, USA (2005). ISBN 1-59593-135-X. doi: 10.1145/1081870.1081918. URL `http://doi.acm.org/10.1145/1081870.1081918`.

[11] Isabelle Guyon/ Jason Weston/ Stephen Barnhill/ and Vladimir Vapnik. *Gene Selection for Cancer Classification Using Support Vector Machines. Mach. Learn.*, 46(1-3):389–422 (March 2002). ISSN 0885-6125. doi:10.1023/A:1012487302797. URL `http://dx.doi.org/10.1023/A:1012487302797`.

[12] Jesse Hoey. *The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way Chi Squared Test. arXiv* (2012). URL `http://arxiv.org/pdf/1206.4881v2.pdf`.

[13] John Joseph Horton and Lydia B. Chilton. *The Labor Economics of Paid Crowdsourcing.* In *Proceedings of the 11th ACM Conference on Electronic Commerce*, EC '10, pages 209–218. ACM, New York, NY, USA (2010). ISBN 978-1-60558-822-3. doi:10.1145/1807342.1807376. URL `http://doi.acm.org/10.1145/1807342.1807376`.

[14] Panagiotis G. Ipeirotis. *Analyzing the Amazon Mechanical Turk Marketplace. XRDS*, 17(2):16–21 (December 2010). ISSN 1528-4972. doi:10.1145/1869086.1869094. URL `http://doi.acm.org/10.1145/1869086.1869094`.

[15] Aniket Kittur/ Susheel Khamkar/ Paul André/ and Robert Kraut. *CrowdWeaver: Visually Managing Complex Crowd Work.* In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1033–1036. ACM, New York, NY, USA (2012). ISBN 978-1-4503-1086-4. doi:10.1145/2145204.2145357. URL `http://doi.acm.org/10.1145/2145204.2145357`.

[16] Aniket Kittur/ Boris Smus/ Susheel Khamkar/ and Robert E. Kraut. *CrowdForge: Crowdsourcing Complex Work.* In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 43–52. ACM, New York, NY, USA (2011). ISBN 978-1-4503-0716-1. doi:10.1145/2047196.2047202. URL `http://doi.acm.org/10.1145/2047196.2047202`.

[17] Anand Kulkarni/ Matthew Can/ and Björn Hartmann. *Collaboratively Crowdsourcing Workflows with Turkomatic.* In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1003–1012. ACM, New York, NY, USA (2012). ISBN 978-1-4503-1086-4. doi:10.1145/2145204.2145354. URL `http://doi.acm.org/10.1145/2145204.2145354`.

[18] Greg Little/ Lydia B. Chilton/ Max Goldman/ and Robert C. Miller. *TurKit: Human Computation Algorithms on Mechanical Turk.* In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 57–66. ACM, New York, NY, USA (2010). ISBN 978-1-4503-0271-5. doi:10.1145/1866029.1866040. URL `http://doi.acm.org/10.1145/1866029.1866040`.

[19] Winter Mason and Duncan J. Watts. *Financial Incentives and the "Performance of Crowds". SIGKDD Explor. Newsl.*, 11(2):100–108 (May 2010). ISSN 1931-0145. doi:10.1145/1809400.1809422. URL `http://doi.acm.org/10.1145/1809400.1809422`.

[20] Jon Noronha/ Eric Hysen/ Haoqi Zhang/ and Krzysztof Z. Gajos. *Platemate: Crowd-sourcing Nutritional Analysis from Food Photographs*. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 1–12. ACM, New York, NY, USA (2011). ISBN 978-1-4503-0716-1. doi:10.1145/2047196.2047198. URL `http://doi.acm.org/10.1145/2047196.2047198`.

[21] Joel Ross/ Lilly Irani/ M. Six Silberman/ Andrew Zaldivar/ and Bill Tomlinson. *Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk*. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 2863–2872. ACM, New York, NY, USA (2010). ISBN 978-1-60558-930-5. doi:10.1145/1753846.1753873. URL `http://doi.acm.org/10.1145/1753846.1753873`.

[22] Luis von Ahn and Laura Dabbish. *Labeling Images with a Computer Game*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326. ACM, New York, NY, USA (2004). ISBN 1-58113-702-8. doi:10.1145/985692.985733. URL `http://doi.acm.org/10.1145/985692.985733`.

[23] Luis von Ahn/ Benjamin Maurer/ Colin McMillen/ David Abraham/ and Manuel Blum. *reCAPTCHA: Human-Based Character Recognition via Web Security Measures*. *Science*, 321(5895):1465–1468 (2008). doi:10.1126/science.1160379. URL `http://www.sciencemag.org/content/321/5895/1465.abstract`.

[24] Frank Wilcoxon. *Individual comparisons by ranking methods. Biometrics Bulletin*, 1(6):80–83 (12 1945). ISSN 00994987.

[25] Tingxin Yan/ Vikas Kumar/ and Deepak Ganesan. *CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones*. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, pages 77–90. ACM, New York, NY, USA (2010). ISBN 978-1-60558-985-5. doi:10.1145/1814433.1814443. URL `http://doi.acm.org/10.1145/1814433.1814443`.

[26] Lixiu Yu/ Paul André/ Aniket Kittur/ and Robert Kraut. *A Comparison of Social, Learning, and Financial Strategies on Crowd Engagement and Output Quality*. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 967–978. ACM, New York, NY, USA (2014). ISBN 978-1-4503-2540-0. doi:10.1145/2531602.2531729. URL `http://doi.acm.org/10.1145/2531602.2531729`.

Appendix A

Ground Truth

A.1 Basic Items

ID	Image 1	Image 2	Image 3
1			
	Title	NIB 45 EURO $$ PUMA SPORT WRISTWATCH SWISS WATCH MOVEMENT LOVE+FOOTBALL	
	Description	ITEM IS BRAND NEW, FROM THE FIFA WORLD (MUNDIAL) FOOTBALL GAMES. WATCH HAS THREE INTERCHANGEABLE TOP COVER , EACH ONE REPRESENTING THE FLAG OF A TEAM PLAYING AT THE FIFA WORLD (MUNDIAL) FOOTBALL GAME PLUS ONE BLACK COVER IF YOU DON'T WISH TO WEAR THE FLAG COLORS INCLUDED IN THE PACKAGE AS SHOWN IN MY PICTURES. ITEM IS BRAND NEW NEVER USED WITH ORIGINAL BOX PAPERS AND INSTRUCTION ON HOW TO USE THIS WATCH.GREAT GIFT IDEA OR GREAT WATCH FOR THE SPORT LOVERS.ITEM COMES WITH WARRANTY FOR 90 DAYS FROM US AND MANUFACTURER WARRANTY OF 2 YEARS IS INCLUDED IN THE BOX.INSTRUCTIONS ARE IN DUTCH,ENGLISH,FRENCH,ITALIAN, CZECH,GERMAN, PORTUGUESE,SPANISH,HUNGARIAN,CHINESE AND JAPANESE.	
	Category	Jewelry & Watches:Watches:Wristwatches	
	Condition	New with tags	
	Price	4.99	

ID	Image 1	Image 2	Image 3
2			
	Title	Apple iPhone 4 - 16GB - Black (Unlocked) Smartphone	
	Description	16GB Black iPhone 4, unlocked by carrier. This was an AT&T phone so it is GSM, can be used internationally. This phone was manufacturer refurbished and then only used for about a week, so it is basically in perfect condition. Includes original packaging, 30-pin USB connector and charger.	
	Category	Cell Phones & Accessories:Cell Phones & Smartphones	
	Condition	Used	
	Price	185.0	

ID	Image 1	Image 2	Image 3
3			
	Title	Lot of (13) 2013 Bowman Sterling Autograph Auto Relic Jersey Games Used	
	Description	This is for a 2013 Bowman Sterling Lot of 13 Game Used Relics and Autos. You get the exact cards that you see in the pictures. PLEASE PAY BY PAYPAL WITHIN 24 HOURS OF AUCTIONS END OR ITEM WILL BE RELISTED. S+H IS 3.99 WITH DELIVERY CONFIRMATION PLEASE CHECK OUT MY OTHER AUCTIONS	
	Category	Sports Mem, Cards & Fan Shop:Cards:Football	
	Condition	Brand New	
	Price	27.0	

ID	Image 1	Image 2	Image 3

4		
	Title	Nespresso Aeroccino Plus & Citiz Coffee Machine Red
	Description	Nespeesso Aeroccino Plus & Citiz Coffee Machine Fully automatic brewing and milk frothing in two sleek, compact units. Works exclusively with Nespresso's premium coffee capsules, which are easy to order for delivery within two business days (for details, visit www.nespresso.com). Innovative Thermoblock technology with stainless-steel heating element guarantees precise temperature control. A 19-bar pressure pump ensures maximum extraction of flavor. Adjustable tray accommodates cups of various sizes (from small mug to travel cup). Removable water tank for easy refilling. Energy-save mode gradually reduces power if unit is left on. Includes Aeroccino Plus milk frother, which quickly heats milk for consistently perfect foam. Frother has two whisk attachments and an auto shutoff feature. Espresso maker: ABS plastic housing. 14 1/2" x 5" x 11" high. 34-fl.-oz.-cap. water tank. 10 lb. 1200W. Milk frother: Stainless-steel and plastic construction. 4" diam., 6-3/4" high. 8-oz. cap. 550W. This product is intended for use in the United States and Canada and is built to United States electrical standards. Posted with eBay Mobile
	Category	Home & Garden:Kitchen, Dining & Bar:Small Kitchen Appliances:Coffee & Tea Makers:Espresso Machines
	Condition	New
	Price	201.0

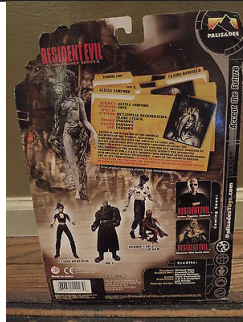| ID	Image 1	Image 2	Image 3
5			
	Title	Nike Mercurial Vapor IX FG - Soccer Shoes Cleats - Metallic Platinum	

	Description	This is a pair of used Nike Vapor IX. They come with the string bag. In overall good condition with some signs of use. Clean and no smells. Mens size 7.5. Shipping is $10.00 and includes tracking. I accept PayPal for payment.	
	Category	Sporting Goods:Team Sports:Soccer:Clothing, Shoes & Accessories:Shoes & Cleats:Men	
	Condition	Pre-owned	
	Price	76.99	

ID	Image 1	Image 2	Image 3
6			
	Title	RARE Series 2 Palisades Resident Evil Code Veronica Alexia Action Figure	
	Description	This RARE and HARD TO FIND action figure will make and AWESOME collectable for any Resident Evil fan! This specific figure is part of the Resident Evil Code Veronica series. Alexia comes complete with Wings, Tail and Alternate Head to Transform into Alexia III and Logo Base. Great item for any RE fan!!! This item is still in its original packaging, unopened and unused. There is very slight wear around the cardboard edging from years of storage and a little adhesive residue on the plastic, most likely from a price sticker. Overall this item is in excellent condition!	
	Category	Toys & Hobbies:Action Figures:TV, Movie & Video Games	
	Condition	New	
	Price	90.0	

ID	Image 1	Image 2	Image 3
7			
	Title	Black Coach purse leather GUC serial number H050-9247	
	Description	Pre-owned Black Coach hobo purse. GUC just because I did use it a couple of times. No stains,marks, or tears. Great condition!!!	
	Category	Clothing, Shoes & Accessories:Women's Handbags & Bags:Handbags & Purses	
	Condition	Pre-owned	
	Price	35.0	

Table A.1: Ground truth for pure crowdsourcing

A.2 Non-branded Item

Image 1	Image 2	Image 3
Title	Saarinen Round Dining Table 47" White Laminate White Base KNOLL DWR	
Description	Living Dining Outdoor Workspace Lighting Floor Accessories MidCentury Modern Saarinen Round Dining Table 47" (White Laminate/White Base) You are bidding on an AUTHENTIC Saarinen Round Dining Table, 47" in White Laminate with White base. Great condition, slight signs of handling. In a 1956 cover story in Time magazine, Eero Saarinen said he was designing a collection to "clear up the slum of legs in the U.S. home." Later that year, he completed his Pedestal Table and TulipTM Chair Collection (1956) and obliterated the "slum" by creating a cast aluminum base inspired by a drop of high-viscosity liquid. This table is manufactured by Knoll according to the original and exacting specifications of the designer. Made in Italy. Pedestal Tables come in a variety of table top materials (veneer, marble and laminate) and base colors (black, white and platinum).The base has an abrasion-resistant Rilsan finish.Each piece is stamped with the KnollStudio logo and Eero Saarinen's signature.Measurements: Assembled Table H 28.25" Diameter 47" Materials: Cast aluminum base with Rilsan finish; bevel-edged top in laminate Laminate: MDF with laminate.	
Category	Home & Garden:Furniture:Tables	
Condition	Used	
Price	1'277.00	

Table A.2: Ground truth non-branded item

Appendix B

Crowdsourcing

B.1   Commission

Ground truth number	End price (USD)	Task	Percentage	Bonus (USD)	Worker ID
1	4.99	Title (Finding)	0.25%	0.01	A3HE1W5T6QO03X
1	4.99	Title (Voting)	0.1%	0.01	A3N7O1NOBGX6U7
1	4.99	Title (Voting)	0.1%	0.01	A1DK26QAO4OOMQ
1	4.99	Description (Improving)	1%	0.05	A2Y9ZNZ0F24GHB
1	4.99	Description (Voting)	0.05%	0.01	A2FF8HA1OWKS83
1	4.99	Description (Voting)	0.05%	0.01	AJAOE1PSNKGUE
1	4.99	Category	0.25%	0.01	A2ZT4MTMEVSLB9
1	4.99	Category	0.25%	0.01	A220ED0LJITW5I
1	4.99	Category	0.25%	0.01	A2V8WJXA0USMZ
1	4.99	Price	0.5%	0.02	A3L99RGPK6FZGH
3	27	Title (Finding)	0.25%	0.06	A23BCMQN9ZU97B
3	27	Title (Voting)	0.1%	0.02	A3N7O1NOBGX6U7
3	27	Title (Voting)	0.1%	0.02	A3I4BYP4DUC475
3	27	Description (Improving)	1%	0.27	A1IA4CST74I1Q8
3	27	Description (Voting)	0.05%	0.01	A3K77RSYXLLUQL
3	27	Description (Voting)	0.05%	0.01	A25F7BNXEN8I5X
3	27	Category	0.25%	0.06	A2ZT4MTMEVSLB9
3	27	Category	0.25%	0.06	A220ED0LJITW5I
3	27	Category	0.25%	0.06	A2V8WJXA0USMZ
3	27	Price	0.5%	0.13	A3L99RGPK6FZGH

Table B.1: Commission results

Appendix C

Machine Learning

C.1 Parameters

The parameters of the classification (Table C.1) and regression (Table C.2) algorithms are part of this chapter.

	kNN	Random Forest	SVC	
	k	Estimators	Kernel	C
iPhone	20	250	Linear	1
Mustang	20	750	Linear	1
Playstation	3	250	Linear	1

Table C.1: Parameters of the classification algorithms

	kNN	Random Forest	SVR	
	k	Estimators	Kernel	C
iPhone	7	2'000	RBF	5'000
Mustang	20	1'000	Linear	100
Playstation	5	500	RBF	10'000

Table C.2: Parameters of the regression algorithms

C.2 Results

The section contains the results of the implemented machine learning approach.

C.2.1 Significance

C.2.1.1 Classification

The results of the significance tests for the iPhone (Table C.3), Mustang (Table C.4) and Playstation (Table C.5) categories.

	kNN	Random Forest	SVC
kNN		3.5875566926347161e-13	1.5291919089303821e-08
Random Forest	3.5875566926347161e-13		4.0439981305274486e-27
SVC	1.5291919089303821e-08	4.0439981305274486e-27	

Table C.3: Significance results classification (iPhone)

	kNN	Random Forest	SVC
kNN		3.789749104003927e-09	1.9962596131950446e-13
Random Forest	3.789749104003927e-09		2.2049869827521087e-22
SVC	1.9962596131950446e-13	2.2049869827521087e-22	

Table C.4: Significance results classification (Mustang)

C.2.1.2 Regression

The results of the significance tests for the iPhone (Table C.6), Mustang (Table C.7) and Playstation (Table C.8) categories.

C.2.2 Classification

The accuracy (Table C.9) and the mean absolut error (Table C.10) of the classification algorithms are illustrated in this subsection. The normalised confusion matrices of the Mustang (Figure C.2) and Playstation (Figure C.4) category are provided too.

C.2.3 Regression

The root mean squared error is summerised in the table C.11. The true and the predicted values were compared in the figures C.1 (Mustang) and C.3 (Playstation).
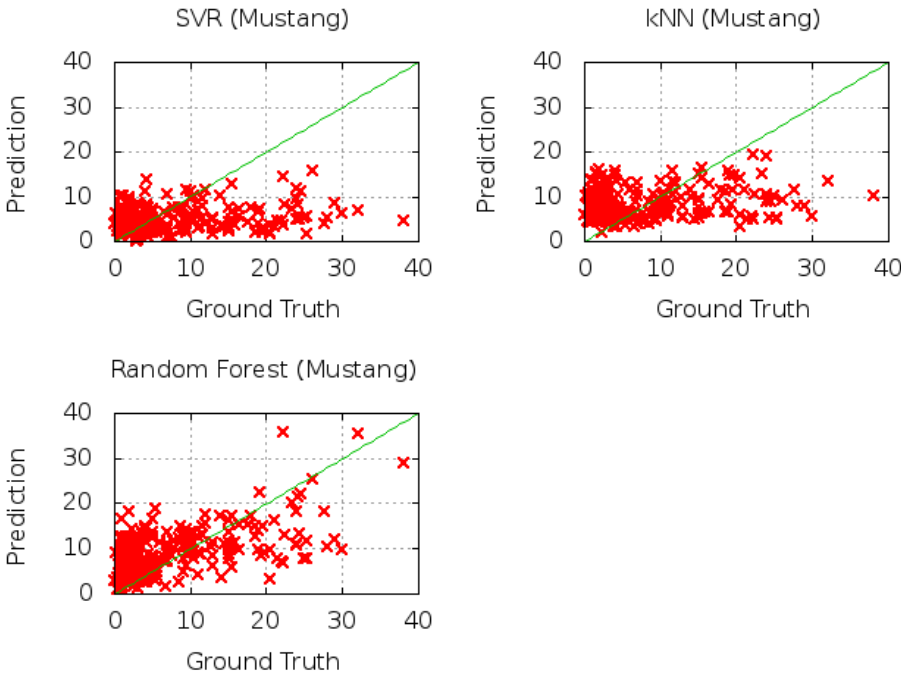
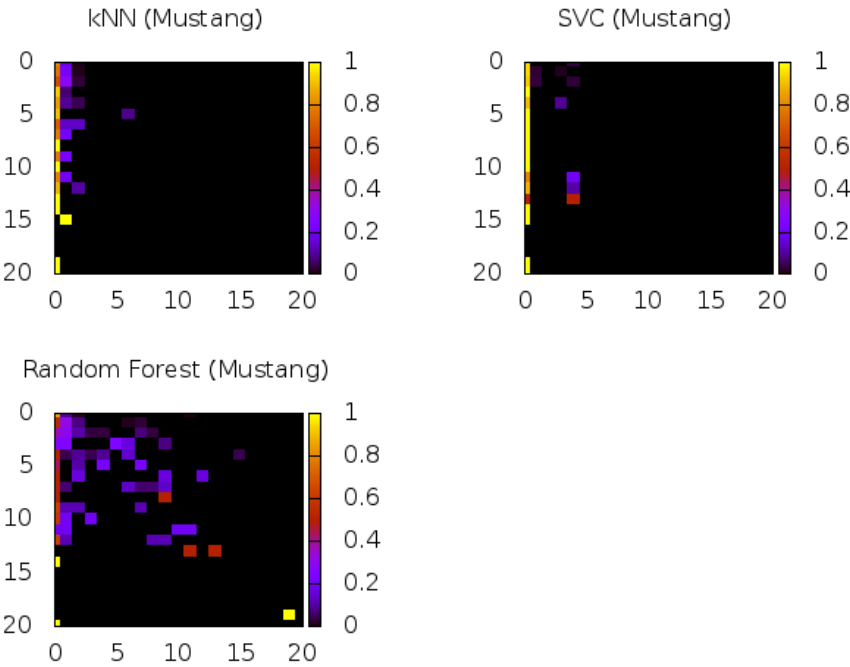Figure C.1: Scatter plot true vs. prediction (Mustang)

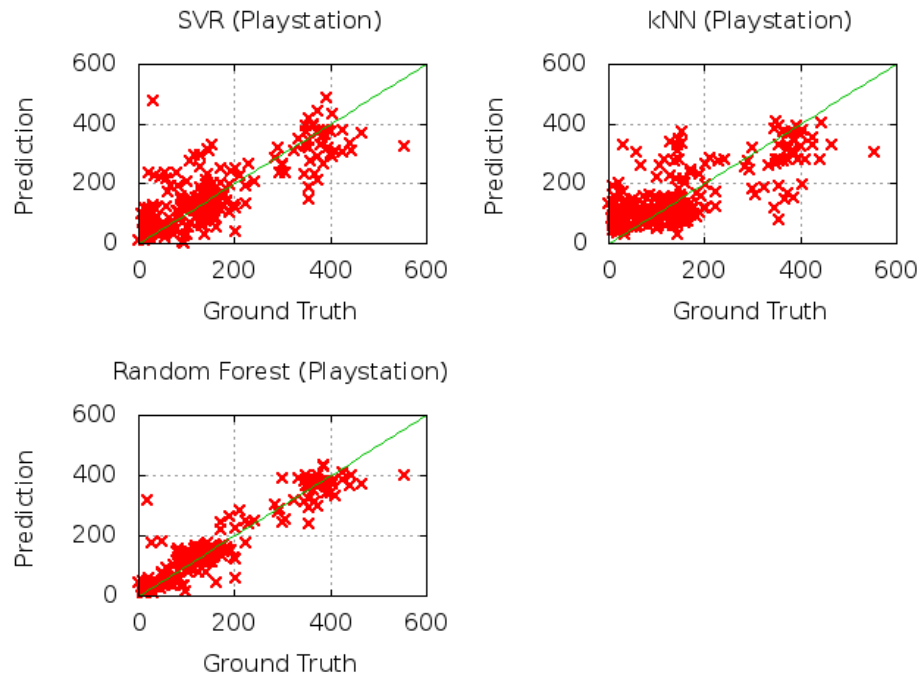Figure C.2: Normalised confusion matrix (Mustang)

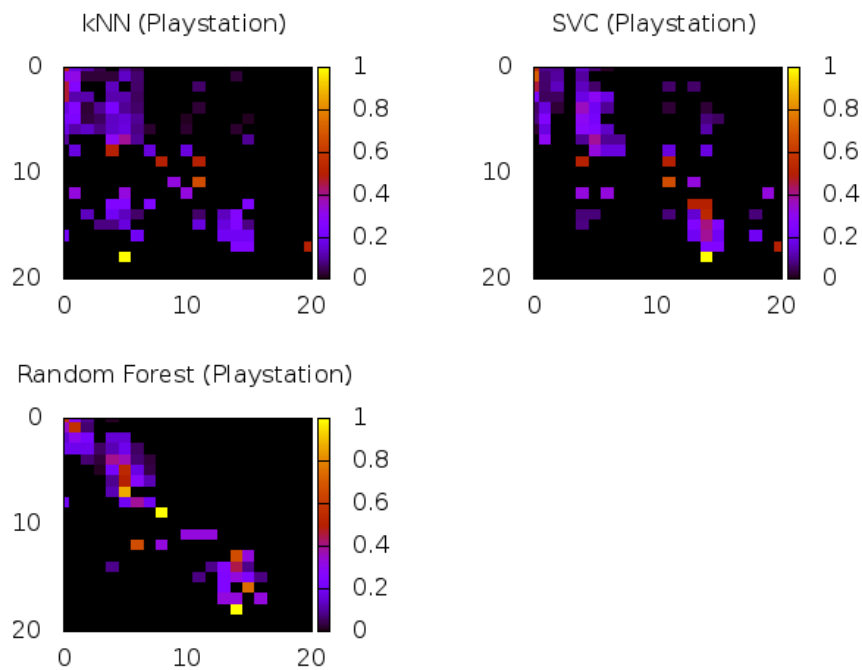Figure C.3: Scatter plot true vs. prediction (Playstation)

Figure C.4: Normalised confusion matrix (Playstation)

	kNN	Random Forest	SVC
kNN		0.0011378352108250606	6.14720172302805e-05
Random Forest	0.0011378352108250606		9.1242838379781361e-05
SVC	6.14720172302805e-05	9.1242838379781361e-05	

Table C.5: Significance results classification (Playstation)

	kNN	Random Forest	SVR
kNN		*0.6311016296123364*	0.22423142184205869
Random Forest	*0.6311016296123364*		0.026965841520582776
SVR	0.22423142184205869	0.026965841520582776	

Table C.6: Significance results regression (iPhone)

	kNN	Random Forest	SVR
kNN		0.048095801005667009	2.4893228869100509e-38
Random Forest	0.048095801005667009		1.2556423582856809e-36
SVR	2.4893228869100509e-38	1.2556423582856809e-36	

Table C.7: Significance results regression (Mustang)

	kNN	Random Forest	SVR
kNN		*0.80679443838141207*	*0.83048173596412633*
Random Forest	*0.80679443838141207*		*0.17292064202909285*
SVR	*0.83048173596412633*	*0.17292064202909285*	

Table C.8: Significance results regression (Playstation)

	kNN	Random Forest	SVC
iPhone	0.10869565217391304	0.19806763285024154	0.16376811594202897
Mustang	0.30281690140845069	0.37323943661971831	0.30633802816901406
Playstation	0.19787985865724381	0.35924617196702008	0.24734982332155478

Table C.9: Results accuracy

	kNN	Random Forest	SVC
iPhone	3.7565217391304349	2.0637681159420289	3.0768115942028986
Mustang	3.26056338028169	2.5387323943661975	3.3450704225352115
Playstation	2.872791519434629	1.1519434628975265	2.4346289752650176

Table C.10: Results mean absolute error

	kNN	Random Forest	SVR
iPhone	92.222790243158784	61.745689964975874	64.313090223009226
Mustang	9.1321357456187897	7.1189517892392047	9.507750986406716
Playstation	82.404952361812889	39.15280385453967	72.629623366881589

Table C.11: Results RMSE