Bremen Big Data Challenge 2017: Predicting University Cafeteria Load

Jochen Weiner, Lorenz Diener, Simon Stelter, Eike Externest, Sebastian Kühl
Christian Herff, Felix Putze, Timo Schulze, Mazen Salous, Hui Liu,
Dennis Küster, and Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany,
`firstname.lastname@uni-bremen.de`

Abstract. Big data is a hot topic in research and industry. The availability of data has never been as high as it is now. Making good use of the data is a challenging research topic in all aspects of industry and society. The Bremen Big Data Challenge invites students to dig deep into big data. In this yearly event students are challenged to use the month of March to analyze a big dataset and use the knowledge they gained to answer a question. In this year's Bremen Big Data Challenge students were challenged to predict the load of the university cafeteria from the load of past years. The best of 24 teams predicted the load with a root mean squared error of 8.6 receipts issued in five minutes, with a fusion system based on the top 5 entries achieving an even better result of 8.28.

Keywords: big data, data analysis, data challenge

1 Motivation

Technical advances in mobile devices and Internet technology, a growing range of devices connected to the Internet, and the general public's readiness to share data and personal information produce more and more data every day. This data accumulates to enormous datasets many of which are collections of unstructured data [3]. Datasets of this size can no longer be handled by trivial means. Instead they require specialist approaches to data generation, data acquisition, data storage and data analysis. These approaches for big datasets call for engineers and data scientists with expertise in data mining, machine learning and data analysis. Using these skills they can leverage big data, uncover latent knowledge hidden in the data and use the unstructured wealth of information to solve problems and extract answers to relevant questions.

We live in an "era of big data" [3]. Using large datasets and the appropriate computing power that is now available we can perform analyses which have not been possible in the past. The usage and analysis of this data will open new possibilities in research, both academic and industrial, and all aspects involved in running a company. If encouraged and nurtured the availability of big data and knowledge in data analysis will drive future development of services, devices and the so-called "Internet of things".

To advance development of techniques applicable to the analysis of such datasets and to gain better understanding of specific sets of data, it has become common practice to hold competitions: Teams can make predictions about a dataset based on given training data and compete against each other to build the system with the best performance on held-out test data (usually not published until after the end of the competition), on platforms specifically built for such competitions [2].

A subcategory of these challenges is those organized specifically for students. Such competitions encourage the participants to familiarize themselves with techniques commonly used in data science and machine learning and allow them to compete specifically against their peers on a reasonably level playing field. One noteworthy competition is the Data Mining Cup, an international student competition with teams from over universities in over 20 countries [4].

2 The Bremen Big Data Challenge

The Bremen Big Data Challenge (BBDC) is a student challenge in the field of big data. This yearly event aims at sparking interest in data science among students. Each year the students are presented with a new big dataset and a task to solve on the dataset. As researchers with an interest in data analysis and lecturers tasked to prepare students for jobs in computer science we are very keen to spread our fascination of data and its analysis to students. Through the challenge as well as through our regular teaching we hope to show the students the diversity of tasks in big data and find talented students willing to take part in the variety of big data-related research at the University of Bremen.

The BBDC was created in 2016 and is open to all students in the federal state of Bremen. Interested students can sign up for a newsletter which keeps them updated with the latest BBDC news, and form teams of one to three participants. On the first of March we publish the big dataset and the corresponding task which is to be completed within 31 days. At the end of the challenge the best five teams are awarded monetary prizes. The data and the reference of the task are then published on the BBDC website [1].

In 2016 the task was to predict players' performance in an online game based on a history of past matches. In 2017 the task was to predict the load in Studentenwerk Bremen's cafeteria [8] on the campus of the University of Bremen.

Compared to other data analysis competitions, the BBDC is a competition that is limited to students local to the Bremen universities only. The task is chosen to allow even those with absolutely no prior experience in machine learning to participate, and the 2017 task directly relates to the lives of University of Bremen students.

Solution submissions for the BBDC 2017 are handled using an automated system that lets each team evaluate their solution on the test data three times a week (a total of 15 submissions which can be saved up). Each team that has submitted at least one solution can also see the leaderboard showing the currently best solution of each participating team.

3 The Bremen Big Data Challenge 2017

For the Bremen Big Data Challenge 2017 we cooperated with the Studentenwerk Bremen [8] which, among other services for students, operates the cafeterias on the campus of the University of Bremen. The question preceding this challenge was: Can we predict the load the main cafeteria on campus will have at a specific time? For this task we counted the number of receipts issued in every five-minute-slot from January 2009 to November 2016. The participants were given the data of these five-minute-slots for 2009 through 2015 and were asked to predict the numbers for the year 2016. The challenge is evaluated using the root mean squared error (RMSE) between the actual cafeteria load in the five-minute-slots and the load prediction.

With a very straight-forward analysis we can draw first conclusions about the data: There is no load around the beginning and end of each year. This is the time the cafeteria is closed between Christmas and the new year. The term in the winter semester usually runs from October to February and from April to July in the summer semester. From personal experience we know that the cafeteria is generally less crowded during the break (February to April and July to October) than during the term. And indeed, the data shows a much higher load during term than during break.

The data covers the whole time period from January 2009 through November 2016. This means that times at which the cafeteria did not issue any receipts (e.g. during the night) are also included. The whole period contains a total of 829,191 five-minute-slots. In 752,841 slots (90.8%) there was no load and there were receipts in 76,350 five-minute-slots (9.2%).

In the average working week all days have a similar distribution of the loads with a small load in the morning and a large peak around midday. The highest load occurs every day at the beginning of lunch time. The total load on an average day varies with considerably less receipt being issued on Fridays than on the other days. On an average day we observe the main load between 11:30 and 14:00 when lunch is served. During this period the load has several peaks, all of which occur five to ten minutes after a full or half hour. Our interpretation of this structure is that people meet for lunch at a full or half hour when their classes end and arrive at the cafeteria's checkout five to ten minutes later.

This first analysis shows that the cafeteria load depends on external information such as the semester times and (obviously) cafeteria closing times. Participants were therefore provided with additional data:

Semester times obviously influence the cafeteria load as described above. The dataset therefore includes the start and end of the term as well as any days without classes and the first-semester orientation week.

Possible cafeteria guests are the group who visit the cafeteria and cause the load there. For this reason the dataset contains the number of students enrolled each semester and the numbers of researchers, lecturers and university staff employed each year.

Weather may influence cafeteria load, since it is possible that fewer people will walk to the cafeteria during a storm or heavy rain. The dataset therefore includes hourly information on wind, rain and temperature, courtesy of Deutscher Wetterdienst [5].

Cafeteria menu varies every day and is available on the cafeteria's website. As some meals are more popular than others, the meals offered on a given day might also influence the load of the cafeteria on that day. The menu is provided as a textual description of each day's menu (the same information available to potential cafeteria visitors).

Participants were explicitly invited to include their own sources of additional information e.g. public holidays, days when the cafeteria was closed, or special events (such as conferences) held at the university.

4 The Course of the Bremen Big Data Challenge 2017

We developed a very straight-forward one-afternoon-baseline before we published the dataset and the task. In this baseline we take into account the whole data from 2009 through 2015 as training data. The system is based on semester times, the day of the week and the time of day. Term and break are handled separately. We calculate the mean of each five-minute-slot per weekday, e.g. the mean of all five-minute-slots starting at 12:00 on a Monday during lecture time. Then we assign this mean to the corresponding time-slots in the test data as the prediction for the year 2016. The prediction of the straight-forward baseline achieves an RMSE of 13.47. When preparing the data and for our baseline system we used the pandas [6] and scikit-lean [7] libraries.

A total of 121 students from all three big universities in Bremen showed interest in the challenge and signed up for the newsletter. 41 students from the University of Bremen and the Hochschule Bremen formed the 24 teams participating in the challenge. Figure 1 shows the progress of the leader's score during the challenge. The first submission with an RMSE of 17.81 was made on the sixth day of the challenge. By the 10th day of the challenge the best score had dropped below an RMSE of 10 and then slowly improved to the final score. The final winning submission with an RMSE of 8.60 was made 5 minutes before the end of the challenge. Table 1 shows the top 5 teams at the end of the challenge.

The number of participants show that students in Bremen are interested in big data and machine learning. Compared to 2016, there was a 10% increase in the number of participants. If the Bremen Big Data Challenge became more popular because of an increased awareness for big data among the students remains to be seen in the next installment of the BBDC in 2018. The results from the BBDC show that as hoped the majority of the teams performed better than our explicitly very straight-forward baseline. The approaches of the first five teams are more complex and more powerful than our baseline, but not all of them used machine learning. This shows that teams could participate and achieve good results without deep knowledge of machine learning. Since this was possible, the teams used substantially different approaches:

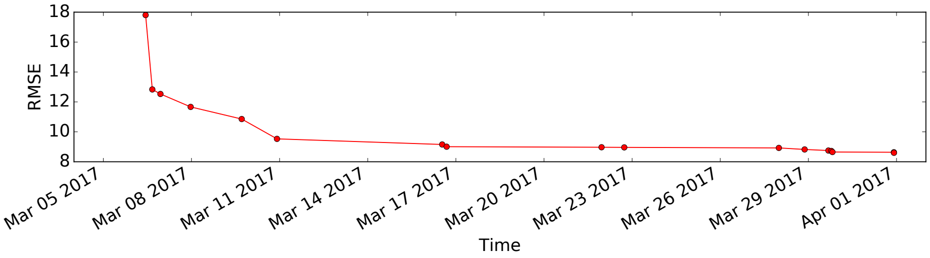

Fig. 1. The development of the leading team's score over time.

Table 1. Results of the challenge: The top 5 and a late-fusion combined system.

		Submission time	
Place	**RMSE**	**First**	**Final**
1	8.60	Mar 15 17:24:47	Mar 31 23:54:49
2	8.63	Mar 12 16:00:58	Mar 31 23:50:55
3	8.75	Mar 06 17:00:49	Mar 29 21:42:20
4	8.80	Mar 06 22:30:16	Mar 31 23:12:45
5	9.21	Mar 11 13:51:54	Mar 31 21:26:22
Fused	8.28		

The winning team considered the customer behaviour underlying the issued receipts: During the term many receipts are issued to students who timed their lunch to be in between lectures, resulting in multiple sale spikes. In contrast, during break students are not restricted by schedules and university staff make up a higher percentage of the customers, leading to a more shallow sales curve. To model this customer behavior they created template sales curves for the summer term, winter term and break: They averaged the receipts issued for each five-minute-slot and then normalized the standard deviation in this average day to one. For each day they predicted a scaling factor for the template sales curve. Multiplying the template sales curve with a predicted scaling factor results in their prediction for that day. The scaling factor was predicted by a regression tree with a maximum depth of 4. It was trained using year, month, day, weekday and semester time as features. The target was the standard deviation of sales (the scaling factor) on the day specified by the feature.

The second-place finisher predicted the receipt count directly using a 2 hidden layer (1000 and 300 ReLU units respectively) feed forward neural network trained using RMSprop (with a learning rate of 0.0005 and a batch size of 128). In addition to the time, they also used weather information as additional features and excluded federal holidays during which the cafeteria was closed. The fourth place finisher used a similar technique (feedforward neural network with slightly different architecture), but did not use weather or holiday information.

While the first- and second place finishers used common machine learning techniques, the team that came in in third place used a rather simpler method for their system: They assigned each slot the average past value according to the mean of past slots with the same time-of-day, day of week, month and part of the academic year after smoothing out the data with a median filter.

The fifth-placing teams entry was special insofar as it was the only highly-placed entry that made use of the cafeteria menu: The system automatically grouped textually similar menus and used them (as well as time and weather features) as input for a gradient boosting regressor.

One similarity between all the winning entries is that they relied heavily on selecting which data to train their systems on: All of them chose to exclude data, sometimes using only the most recent year and often choosing to train their system only for times during which the cafeteria is known to be open and setting all other times to zero (or modeling them separately). While some systems are similar, they do appear to be learning different things: A late fusion of all the winning systems outputs (by taking their mean) results in a new system that outperforms each single system by a large margin.

On average the teams spent 10.5 days between their first and their last submission. None of the first five teams spent less time than average on the challenge and they used all their available submissions. This shows that time and dedication paid off. The best team's prediction missed the number of receipts issued in a five-minute-slot by just 8.6 receipts on average. The system based on the top 5 entries combined results achieved an even better result with an RMSE of 8.28.

5 Conclusion

Today, larger quantities of data and a wider range of data is available than ever before. Working with this data and using this data to answer relevant questions is not an easy task. The Bremen Big Data Challenge aims at sparking interest in data research among students in Bremen. The Bremen Big Data Challenge 2017 focused on the load of the Studentenwerk Bremen's university cafeteria. Participants were supplied with the cafeteria's load in five-minute-slots from 2009 to 2015 and supplementary data such as the cafeteria menu, semester times and weather. Their task was to predict the cafeteria load in the five-minute-slots of the year 2016. 24 teams participated in this year's challenge and achieved a range of good results with the best team's prediction missing the true number of receipts issued in a five-minute-slot by just 8.6 receipts. Combining the top 5 results the number of receipts issued is missed by only 8.28 receipts.

We will continue this tradition with the Bremen Big Data Challenge 2018.

Acknowledgements. We thank the Studentenwerk Bremen for providing us with the dataset for the Bremen Big Data Challenge 2017.

References

1. Bremen Big Data Challenge: `https://bbdc.csl.uni-bremen.de/`, [Online]
2. Carpenter, J.: May the best analyst win (2011)
3. Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mobile Networks and Applications 19(2), 171–209 (2014)
4. Data Mining Cup: `http://www.data-mining-cup.de/en/dmc-wettbewerb/wettbewerb.html`, [Online]
5. Deutscher Wetterdienst – Archiv Monats- und Tageswerte: `http://www.dwd.de/DE/leistungen/klimadatendeutschland/klarchivtagmonat.html`, [Online]
6. McKinney, W.: Data structures for statistical computing in python. In: van der Walt, S., Millman, J. (eds.) Proceedings of the 9th Python in Science Conference. pp. 51 – 56 (2010)
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
8. Studentenwerk Bremen: `http://www.stw-bremen.de/en`, [Online]